UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Miha Batič

PODATKOVNO RUDARJENJE KOT NADGRADNJA PRODAJNE APLIKACIJE ZA POSLOVNO OBVEŠČANJE

DIPLOMSKA NALOGA NA UNIVERZITETNEM ŠTUDIJU

Mentor: doc. dr. Marko Bajec

Ljubljana, 2008

ZAHVALA

Zahvalil bi se svojemu mentorju, doc. dr. Marku Bajcu, za vložen trud, čas in potrpežljivost. Boštjanu Kožuhu se zahvaljujem za konstruktivne pripombe na moje diplomsko delo in vsebinske predloge. Celotnemu kolektivu v Adacti bi se zahvalil za podporo in priložnost za sodelovanje pri zanimivih projektih, kjer sem se veliko naučil. S trenutnim delom so me pripeljali do ideje za to diplomsko nalogo. Zahvalil bi se Sebastjanu Kastelicu za predloge in pripombe pri ustvarjanju diplomske naloge. Urški Novak bi se zahvalil za lektoriranje tega diplomskega dela. Zahvalil bi se tudi strašem, ki so me podpirali na tej dolgi poti do diplomiranja, mi stali ob strani in mi vse skupaj tudi omogočili. Na koncu pa bi se zahvalil še vsem, ki so me podpirali in spodbujali, da bi čim prej diplomiral.

KAZALO

SEZNAM UPORABLJENIH KRATIC IN SIMBOLOV

ERP – Enterprise Resource Planning - Celovito povezana in na poslovnem modelu temelječa sestava uporabniških programov, ki ob uporabi sodobne tehnologije zagotavlja vsem poslovnim procesom organizacije in njenim poslovnim partnerjem možnosti načrtovanja, razporejanja virov in ustvarjanja dodane vrednosti

BI – Business Intelligence - poslovno obveščanje. Sistem, ki omogoča analizo podatkov o poslovanju organizacije in posledicah sprejetih odločitev.

CRM – Customer Relationships Management - sistem za obvladovanje razmerij s kupci. Poslovna strategija, ki se usmerja na potrebe strank in z uporabo informacijske tehnologije zbira in v ta namen izkorišča vse pomembne informacije

DW – Data Warehouse – podatkovno skladišče: podatkovna zbirka, kjer so shranjeni podatki iz različnih podatkovnih virov.

OLAP – On-Line Analytical Processing – tehnika za analiziranje s funkcionalnostmi, kot so npr.: sumarizacija, konsolidacija, agregacija in tudi ogled podatkov z različnih kotov.

SQL – Structured Query Language – strukturiran povpraševalni jezik za delo s podatkovnimi bazami.

ROI – Return On Investment – kazalec, ki opisuje donosnost naložbe.

MDX - Multidimensional Expressions – poizvedovalni jezik za prebiranje podatkov iz OLAP kock.

NLP – Natural language processing – tehnika, ki se uporablja pri podatkovnem rudarjenju po tekstu.

IR – Information retrieval – tehnika, ki se uporablja pri podatkovnem rudarjenju po tekstu.

HTML – Hyper Text Markup Language – programski jezik, ki omogoča izdelavo spletnih strani.

XML – eXtensible Markup Language – standard, ki služi definiciji podatkovnih struktur. Omogoča zelo kompleksne strukture dokumentov in podatkov.

KPI – Key Performance Indicator **–** Ključni kazalnik poslovanja. Kazalnik, s katerim je mogoče pri pripravi dogovora določiti ciljne vrednosti elementov poslovanja.

POVZETEK

V današnjem svetu, ko poslovanje podjetja podpirajo kompleksni informacijski sistemi, se soočamo z veliko količino podatkov in le-ta narašča na dnevni ravni. Brez dobrega poznavanja podatkov, ki so zapisani v podatkovnih bazah različnih informacijskih sistemov, ki jih podjetje uporablja, so ti podatki ničvredni.

V tem diplomskem delu sem razdelal algoritme podatkovnega rudarjenja na operacije in jih opisal. Že pri opisovanju algoritmov in skupin sem naletel na veliko različnih razdelitev. Izmed vseh interpretacij posameznih skupin sem izbral najbolj primerno razdelitev.

Osredotočil sem se na uporabo tehnik podatkovnega rudarjenja na področju prodaje, ki je ena osnovnih poslovnih funkcij podjetja. Vsi vsaj površno poznamo funkcijo prodaje, saj se z njo srečujemo (skoraj) vsak dan. Opisal sem možnost uporabe podatkovnega rudarjenja po področjih marketinga, upravljanja s tveganji in nadzora nad prevarami. Glede na to, da je marketing ena ključnih dejavnosti v prodaji, sem se najbolj osredotočil na ta del.

V zadnjem delu diplomskega dela sem v orodju Microsoft Analysis Services realiziral nekatere izmed možnih uporab podatkovnega rudarjenja, in sicer za:

- Navzkrižno prodajo, analizo nakupovalne košarice,
- Segmentacijo trga,
- Napovedovanje prodaje,
- Ohranitev kupca.

Ti pogledi so se mi na obstoječih podatkih iz ERP sistema Microsoft Dynamics NAV zdeli najbolj smiselni. Ugotovil sem tudi, da moramo v nekaterih primerih, kljub temu, da Microsoft Analysis Services omogoča povezavo na relacijsko podatkovno bazo, še vedno narediti podatkovno skladišče. To je zelo odvisno od organizacije podatkov v tabelah.

Izkazalo se je, da je zaradi kompleksnosti in zanimivosti prodajne funkcije obravnavanje funkcije prodaje nekoliko preveč obširno. Ravno iz tega razloga sem se omejil pri implementaciji možnih uporab podatkovnega rudarjenja na zgornje štiri.

Ključne besede:
Podatkovno rudarjenje, prodaja, povečanje prodaje, poslovno odločanje, napovedovanje, algoritmi za podatkovno rudarjenje.

ABSTRACT

These days companies have complex information systems with a vast amount of data. Amount of data increases every day. With insufficient knowledge of data, stored in company databases, the data is worthless.

In this diploma I have divided data-mining algorithms into operations and described them. During the description of these algorithms and groups of algorithms I have encountered several different classifications. From among them I have chosen the most suitable one.

I focused on the usage of data-mining techniques in the sales department, which is one of the basic business functions. Everybody knows sales function; we come across the sales function every day. I described possible use of data-mining in the marketing field, risk management and fraud management. I focused mainly on marketing part since it is very important activity.

In the last part of this diploma I realized some of the potential uses of data-mining in Microsoft Analysis Services. I realized:

- Cross-sales, Market basket analysis,
- Segmentation,
- Sales forecast,
- Customer retention.

The data from Microsoft Dynamics NAV ERP system allowed me to perform data-mining. I also discovered, that even though Microsoft Analysis Services gives us the opportunity to perform data-mining only with connecting to relational databases, we still need data warehouse in some cases. This is very much dependent on data organization in data tables.

Due to a complexity and interesting details about sales function I believe that discussing sales function is very extensive. This is also the reason why I limited by only implementing possible uses of data-mining to the above four.

Keywords:
Data-mining, sales, sales improvement, business intelligence, forecasting, data-mining algorithms.

UVOD

Nahajamo se v dobi, ko ima veliko podjetij že postavljeno zaledje, tj. informacijski sistem ali več le-teh, v katerem se nahaja veliko raznovrstnih podatkov. Celovite poslovne rešitve (ERP) so podjetjem omogočile, da poenotijo obdelavo poslovnih dogodkov in na enem mestu zberejo podatke o poslovanju. Seveda podatki sami po sebi nimajo velike vrednosti, če jih ne znamo pravilno interpretirati in preoblikovati v informacije, ki so za podjetje v informacijski dobi ključne.

Količina in raznolikost zbranih podatkov naraščata, hkrati pa se pojavljajo potrebe po primerljivosti s konkurenčnimi okolji in podatki iz javnih zbirk. Podatke, ki jih lahko dobimo iz najrazličnejših virov, je potrebno medsebojno povezati, analizirati in pripraviti poročila, ki so primerna in predvsem uporabna za sprejemanje pravilnih poslovnih odločitev.

Uvedba rešitve za poslovno obveščanje podjetjem to omogoča in s tem zagotavlja, da pri svojem razvoju naredijo korak naprej. Rešitve za poslovno obveščanje so namenjene tako vodilnim v podjetjih, lastnikom kot tudi vodjem oddelkov in nekoliko bolj zahtevnim uporabnikom. Vodilnim pomagajo pri sprejemanju poslovnih odločitev, lastnikom pa omogočajo vpogled v poslovanje in nadziranje svojih naložb.

S sistemi za BI lahko predvsem nadziramo poslovanje in dogajanje v podjetju v preteklosti. Če želimo planirati poslovanje podjetja v prihodnosti, lahko to v določeni meri napovemo s pomočjo različnih tehnik umetne inteligence. V tem diplomskem delu se bom osredotočil na podatkovno rudarjenje.

1 Prodaja kot osnovna poslovna funkcija

Danes si le težko predstavljamo podjetje brez prodaje oz. prodajne funkcije. Zato ni čudno, da za prodajo velja, da je ena izmed osnovnih poslovnih funkcij [5]. Podjetje začne in konča svoje delovanje na trgu. Z raziskavo trga v podjetju ugotavljajo potrebe potrošnikov, ugotovljene potrebe potrošnikov pa določajo načrtovanje ustvarjanja poslovnih učinkov (izdelki oz. storitve). S prodajo preidejo poslovni učinki k potrošniku. V tržnem gospodarstvu lahko že majhne napake v prodaji ovirajo običajni tok ustvarjanja in razpečave poslovnih učinkov, kar posledično pomeni, da podjetje ne dobi pravočasno denarja. Prodajna funkcija zagotavlja pretvarjanje poslovnih učinkov v denar. S prodajo ustvarjamo možnost za nadaljnje poslovanje in razvoj podjetja.
Osnovne naloge prodajne funkcije:

a) Raziskovanje prodajnega trga,
b) Prodaja poslovnih učinkov (pogovori o prodaji, izdelava ponudb in pogodb, …),
c) Skladiščenje in prevoz poslovnih učinkov,
d) Beleženje ali razvid prodaje,
e) Ekonomsko analiziranje in kontrola prodaje,
f) Poslovno oglaševanje,
g) Načrtovanje prodaje.

Naslednje štiri opisane naloge prodajne funkcije sem izbral zato, ker jih je mogoče informacijsko podpreti.

Raziskovanje prodajnega trga omogoča podjetjem, da z dobrim poznavanjem razmer na trgu povečajo urejenost in uspešnost prodajnega poslovanja in s tem pripomorejo k izboljšanju zadovoljevanja potreb potrošnikov. Proučevanje prodaje se ob proučevanju potreb uporabnikov na posameznem področju nanaša predvsem na kupno moč potrošnikov, uveljavljenost blagovnih znamk, zasičenost trga, … Na podlagi proučevanja prodaje lahko izberemo eno izmed naslednjih tehnik za povečevanje dodane vrednosti prodaje, npr: *povečanje količine izdelkov*, ki jih prodamo, ali *povečanje cene izdelkov*, ki jih prodajamo.

Poslovno oglaševanje lahko vidimo kot obveščanje, prepričevanje in pridobivanje potrošnikov za nakup poslovnih učinkov. Usmerimo se lahko na stalne potrošnike in občasne potrošnike (te skušamo na novo pridobiti med svoje potrošnike). Izboljšan marketing pa pripomore k višji prodaji in s tem k večjemu zaslužku podjetja.

V *razvidih prodaje* podjetja vključujejo podatke o prodaji poslovnih učinkov. Razvidi služijo kot podlaga za odločitve in za sprotni nadzor delovanja v prodaji. Pri *ekonomskem analiziranju in kontroli prodaje* skuša podjetje ugotavljati dejansko stanje prodaje in dejavnike, ki na to stanje vplivajo. Kontrola prodaje vsebuje primerjavo doseženih prodajnih rezultatov z načrtovanimi oz. pričakovanimi rezultati. Kontrola prodaje je z vidika uspešnosti poslovanja podjetja pomembna:

- ker se spremembe na trgu najhitreje odražajo prav v prodaji,
- ker lahko napake v procesu ustvarjanja poslovnih učinkov in prodaji vplivajo na poslovanje podjetja,
- ker le uspešna prodaja poslovnih učinkov ustvarja možnosti za razvoj družbe.

Z *načrtovanjem prodaje* podjetje konkretizira svoje prodajne cilje in pričakovane prodajne rezultate. Načrtovanje prodaje se nanaša na ocenjevanje prodajnih možnosti na sedanjih in novih trgih, organizacijo prodaje ter določanje metod in oblik prodaje. Podlaga za načrtovanje prodaje so zlasti podatki raziskave trga poslovnih učinkov pa tudi opravljene analize bodočih tržnih gibanj.

Vidimo lahko, da je zadnje štiri točke smotrno informacijsko podpreti. S tem, ko jih informacijsko podpremo, lahko pričakujemo izboljšanje prodajnih rezultatov, saj je pregled prodaje tako bolj transparenten in omogoča boljši pregled nad poslovanjem. Beleženje ali razvid prodaje in ekonomsko analiziranje in kontrola prodaje sodita bolj v poslovno obveščanje, medtem ko poslovno oglaševanje in načrtovanje prodaje sodita med naloge, kjer bi lahko nekaj več pridobili z uvedbo podatkovnega rudarjenja.

1.1 Kako lahko informatika pripomore k večji prodaji

Kot sem zapisal že zgoraj lahko na povečanje dodane vrednosti lahko gledamo z različnih stališč, in sicer:

a) povečamo lahko količino izdelkov, ki jih prodamo,
b) povečamo lahko cene izdelkov, ki jih prodajamo,
c) izboljšamo kvaliteto izdelkov,
d) izboljšamo lahko marketing,
e) stranki ponudimo komplementarne izdelke oz. storitve.

Obstoječa aplikacija za poslovno obveščanje za analizo prodaje med drugim omogoča pregled prodaje in enostavno planiranje prodaje preko tako imenovane KAJ-ČE (WHAT IF) analize, kjer zadostimo prvima dvema točkama (povečanje količine, povečanje cen). Četrto točko je težje analizirati, učinek marketinga je namreč težko merljiv in na trenutke težje dosegljiv. S podatkovnim rudarjenjem pa lahko povečamo učinkovitost marketinga in navzkrižne prodaje (točka e). Posledično se s tem poveča prodaja.

Z orodjem za poslovno obveščanje in razvojem aplikacije za prodajo se omogoči podjetju spremljanje prodaje in hitra odzivnost na morebitne dogodke na trgu. V zadnjem času pa se vse več podjetij ukvarja tudi z načrtovanjem prodaje, kar zahteva vpeljavo naprednejših algoritmov in tehnik, kot je podatkovno rudarjenje.

Lahko bi rekli, da je poslovno obveščanje kurativa, saj uporabniku omogoča vpogled v poslovanje podjetja v preteklosti. Na podlagi informacij, ki jih vodilni dobijo preko aplikacij za poslovno obveščanje, se lahko v podjetju odločijo kako bodo vodili podjetje v prihodnosti.

Koncept podatkovnega rudarjenja pa bi lahko označili kot preventiva, saj nam s pomočjo nekoliko zahtevnejših algoritmov pomaga pri »napovedovanju prihodnosti«.

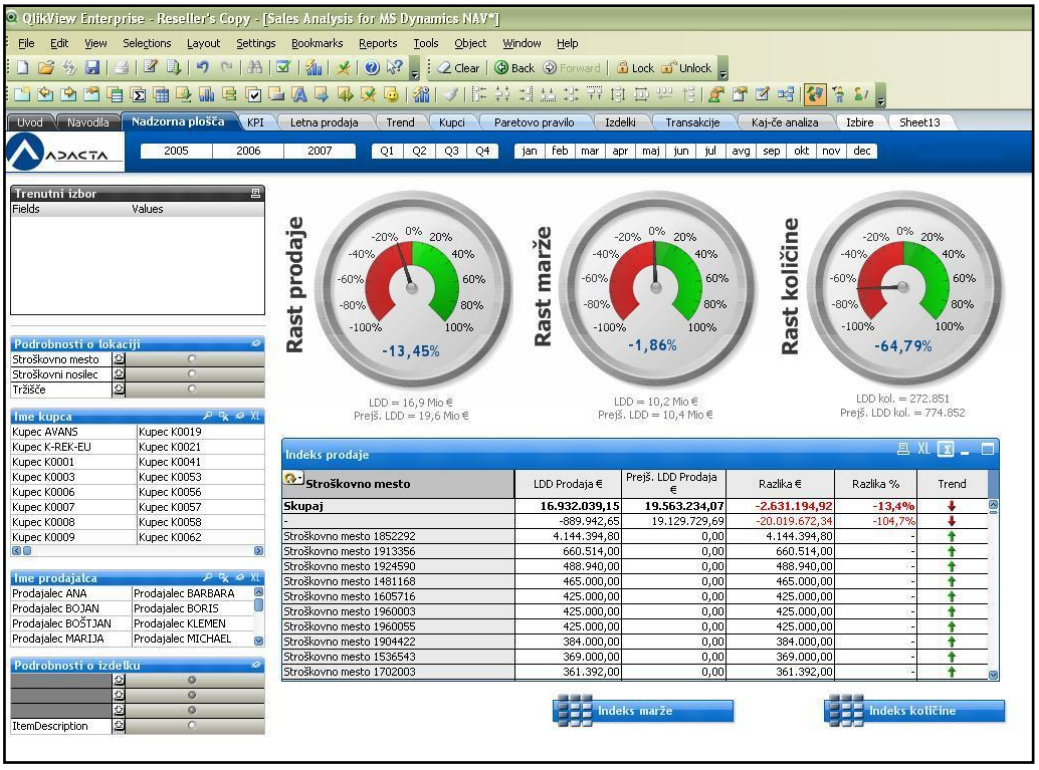
1.2 Obstoječa aplikacija za poslovno obveščanje: Prodaja

Aplikacija za analizo prodajo je ena prvih ključnih aplikacij, ki jih zahteva naročnik, ko se odloči za sistem za poslovno obveščanje.

Aplikacija omogoča prikaz različnih ključnih kazalcev uspeha, kot so rast prodaje, rast marže, rast količine in indeks prodaje (prodaja v prejšnjem obdobju v primerjavi s trenutno prodajo). Poleg časovne primerjave omogoča tudi prikazovanje trenda prodaje, marže in količine v izbranem obdobju ter najboljše prodajalce in kupce. Pri večini prikazov omogoča vrtanje v globino do posameznega računa. Aplikacija omogoča tudi Kaj-Če analizo. S tem ko z drsnikom uravnavamo pričakovano ceno, strošek ali količino, se v preglednici sproti preračunavajo nove vrednosti.

Prikaz podatkov o prodaji v standardni aplikaciji omogoča pregled po naslednjih dimenzijah:

- datum knjiženja (tudi leto, mesec, mesec-leto, četrtletje)
- izdelek
- kupec
- kategorija izdelka
- prodajalec, skrbnik



Slika 1.1. Izgled aplikacije za prodajo – nadzorna plošča

Aplikacija je namenjena uporabnikom Microsoft Dynamics NAV. Microsoft Dynamics NAV je ERP sistem za podjetja.

Podatki o prodaji se v Microsoft Dynamics NAV nahajajo v tabelah:

- Postavka kupca (Cust. Ledger Entry) – podatki o postavkah, vezanih na kupca
- Podrobna postavka kupca (Detailed Cust. Ledg. Entry) – podatki o podrobnih postavkah, vezanih na kupca
- Postavka artikla (Item Ledger Entry) – podatki o postavkah, ki so vezane na artikel
- Postavka vrednosti (Value Entry) – podatki o vrednostih postavk
- Artikel (Item) – podatki o vseh artiklih
- Kupec (Customer)– podatki o kupcih
- Delavec (Employee) – podatki o delavcih

Zadnje tri tabele so predvsem šifranti, ki jih bomo uporabili tudi pri podatkovnem rudarjenju, kjer bomo proučevali določene zakonitosti v povezavi z njimi. V ostalih tabelah pa se nahajajo podatki o poslovnih transakcijah.

2 Podatkovno rudarjenje

Podatkovno rudarjenje oz. (v literaturi se pojavlja tudi kot pojem) iskanje znanja v podatkih je netrivialen proces identificiranja

- veljavnih,
- neobičajnih,
- potencialno uporabnih in
- predvsem razumljivih

vzorcev v podatkih. [3]

Vsak dan v poslovnem sistemu nastane lahko tudi po več deset tisoč transakcij, ki so odraz sodelovanja poslovnega sistema z okoljem in, ne na zadnje, tudi dobrega poslovanja. S tem, ko število transakcij narašča, postaja vse težje iz ogromne količine podatkov izluščiti pomembne informacije in tudi znanje. Informacije in znanje, ki ga pridobimo iz podatkov, lahko uporabimo za vodenje poslovanja, kontrolo proizvodnje, analizo trgov do znanstvenih raziskav, itd. Evolucijska pot pri razvoju podatkovnih baz poteka z razvojem naslednjih funkcionalnosti:

1. **izdelava podatkovne baze in zbirk podatkov**
2. **upravljanje s podatki** (vključuje zbiranje in shranjevanje podatkov, procesiranje transakcij)
3. **analiziranje podatkov in razumevanje le-teh** (vključuje BI, podatkovna skladišča in podatkovno rudarjenje)

Podatki so lahko zbrani na različne načine, eden izmed njih je v obliki podatkovnega skladišča. Tehnologija podatkovnih skladišč vključuje čiščenje podatkov, integracijo podatkov in implementacijo OLAP kocke (OLAP, MOLAP, ROLAP). Kljub temu, da OLAP orodja omogočajo večdimenzionalne analize in sprejemanje odločitev na podlagi velike količine podatkov (**decision making)**, še vedno potrebujemo dodatna orodja za različne analize v globino (**in-depth analysis**), kot so klasifikacija, **razvrščanje v skupine**, in **označevanje podatkov** skozi čas. V zadnjem času se pojavljajo tudi orodja, ki za izdelavo večdimenzionalnih analiz ne potrebujejo OLAP kock. Eno takih orodij je QlikView [10], ki si vse podatke shrani v spomin in jih obdeluje v realnem času v spominu.

Pogosto se dogaja, da uporabniki sprejemajo odločitve na podlagi intuicije, in ne toliko na podlagi uporabnih informacij (ang. information-rich data) iz podatkovnih baz, kot bi (v večini primerov) moralo biti. To se dogaja predvsem zaradi tega, ker odločevalci (ang. **Decision makers**) nimajo na voljo pravih orodij oz. tehnologij, ki bi iz podatkov izluščila dragoceno znanje, in zaradi tega, ker je bilo v preteklosti, ko ni bilo toliko podatkov, to povsem dovolj. Orodja za podatkovno rudarjenje omogočajo analizo podatkov in lahko uporabniku razkrijejo pomembne vzorce v podatkih (ang. Data patterns), kar lahko veliko prinese k poslovnim strategijam in zbirkam znanja.

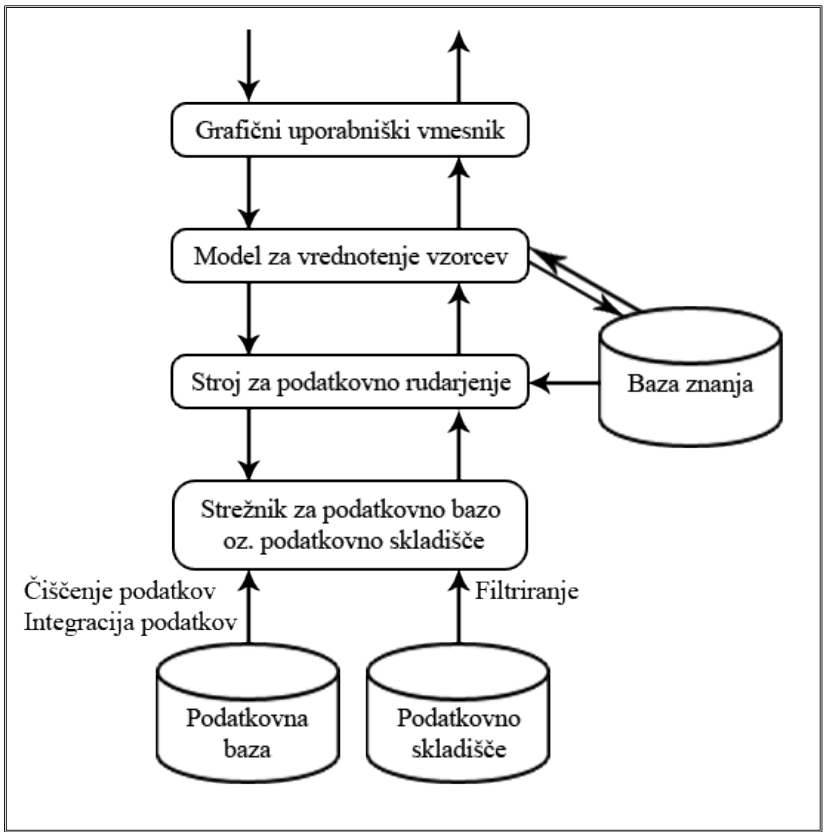
Vse več podjetij se zaveda vrednosti znanja, ki se skriva v podatkih, ki nastajajo vsak dan med izvajanjem poslovnih procesov podjetja. V teh podatkih težko opazimo koristne informacije brez pomoči orodja, ki je temu namenjeno. Zato ni čudno, da je podatkovno rudarjenje zelo perspektivna panoga v računalniški industriji. Kljub sorazmerno visokim stroškom vpeljave, se

vpeljava še vedno obrestuje pri podjetjih, ki imajo opravka z veliko količino podatkov in želijo izvajati pravočasne in pravilne odločitve [2, 11].

2.1 Arhitektura sistema za podatkovno rudarjenje

Arhitektura sistema za podatkovno rudarjenje sestoji iz naslednjih gradnikov:

- **Podatkovna baza, podatkovno skladišče** je ena ali več podatkovnih baz, podatkovnih skladišč, ali ostalih datotek, kjer se nahajajo podatki. Pomembno je, da se nad podatki v teh podatkovnih bazah naredi čiščenje podatkov (ang. **data cleaning**) in integracija podatkov (ang. **data integration**). Kvalitetni podatki so zelo pomembni pri izvedbi podatkovnega rudarjenja!
- **Strežnik za podatkovno bazo oz. podatkovno skladišče** je zadolžen za hranjenje in črpanje podatkov iz podatkovnih baz.
- **Baza znanja** je znanje o problemski domeni, ki ga uporabljamo za iskanje ali vrednotenje za nas zanimivih vzorcev v podatkih.
- **Stroj za podatkovno rudarjenje** je pogoj za izvajanje podatkovnega rudarjenja in vsebuje skupino funkcionalnosti za izvedbo opravil kot so označevanje, asociacije, klasifikacija, razvrščanje v skupine, evolucija, analiza deviacije.
- **Model za vrednotenje vzorcev** je komponenta, ki tipično vsebuje mere zanimivosti in se s pomočjo interakcije z moduli za podatkovno rudarjenje odloča za smer iskanja proti zanimivim vzorcem.
- **Grafični uporabniški vmesnik** služi kot povezava med uporabnikom in sistemom za podatkovno rudarjenje.



Slika 2.1. Arhitektura tipičnega sistema za podatkovno rudarjenje. [4, stran 8]

Algoritmi za podatkovno rudarjenje morajo biti **zmogljivi** in **razširljivi**. Za razširljive algoritme velja, da izvajalni čas narašča linearno oz. proporcionalno z velikostjo podatkovne baze. Razširljivi algoritmi morajo imeti tudi dovolj sistemskih resursov (glavni pomnilnik, prostor na trdem disku, ...) za izvajanje.

2.2 Koncepti in algoritmi podatkovnega rudarjenja

Tehnike podatkovnega rudarjenja lahko povežemo v skupine oz. operacije. Glavne operacije, ki jih poznamo pri podatkovnemu rudarjenju so [1]:

- **Napovedno modeliranje** (ang: Predictive Modeling),
- **Segmentacija podatkov** (ang: Database Segmentation),
- **Analiza povezav** (ang: Link Analysis),
- **Odkrivanje ostopanj** (ang: Deviation Detection).

Napovedno modeliranje je proces soroden človeškemu procesu učenja s pomočjo izkušenj, kjer si s pomočjo opazovanja zgradimo model določenega pojava [1, strani: 64-65].
Primer: otrok v svojem otroštvu opazuje različne pasme psov in na podlagi lastnosti, ki so skupne različnim pasmam, lahko še nepoznane pasme identificira kot pse.

Segmentacija podatkov se običajno uporablja za odkrivanje homogenih podskupin strank v podatkovni bazi, kar omogoča izboljšanje natančnosti profilov strank [1, strani 66-67]. Cilj razvrščanja v skupine je, da so objekti znotraj skupine kar se da podobni oz. homogeni, in hkrati čim bolj različni od ostalih objektov v drugih skupinah. Bolj kot so si objekti znotraj skupine podobni in večja kot je razlika z objekti v drugih skupinah, boljše oz. bolj uporabno je razvščanje v skupine [6].
V literaturi se pogosto omenja tudi izraz razvrščanje v skupine (ang. Cluster Analysis) [4, strani 335-335]. Za razliko od napovednega modeliranja, segmentacija podatkovne baze analizira podatke brez da bi imeli začetno množico učnih podatkov. V splošnem to pomeni, da nimamo začetne učne množice podatkov, ker nam niso znani. Pomembna je tudi testna množica, na kateri preverimo pravilnost pridobljenih rezultatov podatkovnega rudarjenja. Razvrščanje v skupine nam omogoča, da dobimo opise razredov (razrede).
S tem ko podatkovna baza raste je pogosto potrebno segmentirati podatke v skupine podatkov, nato pa nad posamezno skupino podatkov izvedemo operacijo podatkovnega rudarjenja, kot npr. napovedno modeliranje. Razdelitev na podmnožice zelo razbremeni in pohitri postopek podatkovnega rudarjenja.
Algoritem za segmentacijo podatkov deluje brez posredovanja uporabnika glede tega, kakšne segmente uporabnik hoče in koliko segmentov naj najde v podatkovni bazi. S tem se tudi odstranijo predsodki ljudi oziroma intuicija in algoritem s tem poišče resnično naravo vpliva podatkov. Temu rečemo nenadzorovano učenje, tudi neusmerjeno rudarjenje (ang: unsupervised learning).
Primer: podpopulacija je lahko opisana kot »bogati, starejši moški« ali »mestna, izobražena ženska«. Posamezno podpopulacijo lahko nato naslavljamo različno, odvisno od profila (kaj stranka želi).

Analiza povezav je operacija, ki v nasprotju z napovednim modeliranjem in segmentacijo podatkov skuša označiti vsebino podatkovne baze oz. tabele v celoti [1, strani 68-69]. Skuša torej poiskati povezave med posameznimi zapisi oz. skupinami zapisov. Te povezave imenujemo tudi asociacije.
Primer: iskanje asociacij med izdelki oz. storitvami, ki jih stranke običajno kupijo skupaj ali v določenem časovnem zaporedju.

Odkrivanje odstopanj je operacija, ki se je začela uveljavljati v zadnjem času. Večkrat ravno odkrivanje odstopanj pripelje do resničnih spoznanj, saj odstopajoči podatki (ang. outliers**)** pokažejo odstopanje med vnaprej pričakovano lastnostjo in realno [1, stran 69]. Danes se uporablja statistične in vizualizacijske metode za odkrivanje odstopanj.
Primer: predvsem pri odkrivanju prevar pri uporabi kreditnih kartic, zavarovanj.

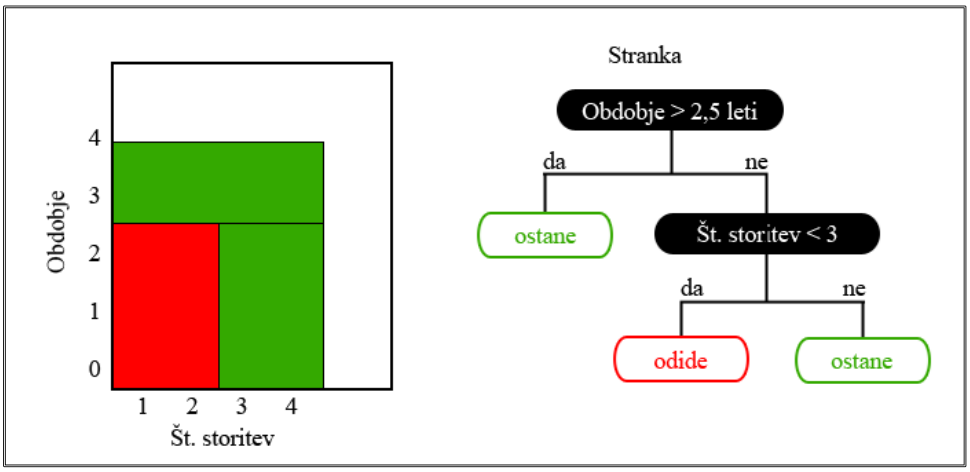
2.3 Tehnike podatkovnega rudarjenja

2.3.1 Napovedno modeliranje

Klasifikacija in predikcija sta dva načina analize podatkov, ki jih uporabimo za izdelavo modelov pomembnih podatkovnih razredov oz. za napovedovanje prihodnjih trendov v podatkih.

Klasifikacija je proces iskanja modela (modelov), ki opisujejo različne razrede in koncepte podatkov. Pod klasifikacijo spadata metodi **indukcije dreves** in **nevronska indukcija**, ki sta obe metodi nadzorovanega učenja (ang: supervised learning). Nadzorovano učenje je proces avtomatične izdelave klasifikacijskega modela iz učne množice. Zapisi v tej učni množici morajo pripadati manjši množici razredov, ki so jih analitiki vnaprej določili, kar pomeni, da poznamo rezultat [1, strani 70-78]. Ko je model enkrat izpeljan, lahko na podlagi izpeljanega modela avtomatično napovemo v kateri razred bo razporejen še nerazporejen razred. Nadzorovana izpeljava razredov je lahko bodisi **nevronska** bodisi **simbolična**. Nevronska izpeljava je prikazana kot slika uteženih povezav med vozlišči, simbolična izpeljava pa kot odločitveno drevo ali kot »**IF THEN**« pravila.

Indukcija dreves zgradi predikcijski model kot odločitveno drevo, pogosto kot binarno odločitveno drevo. Algoritem začne z identifikacijo najbolj pomembne spremenljivke, to je spremenljivka, ki velja za najbolj pomembno pri klasifikaciji.

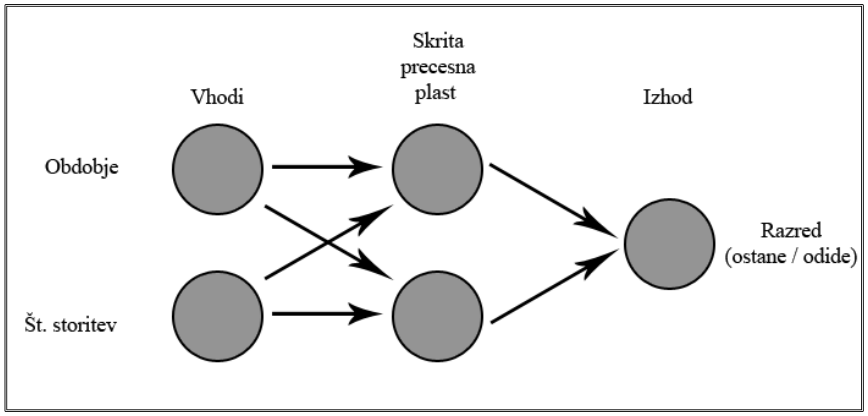


Slika 2.2. Binarno odločitveno drevo.

Indukcija dreves je zelo učinkovita, saj je časovno manj zahtevna, hkrati pa omogoča zelo intuitivno analiziranje rezultatov. Slaba lastnost indukcije dreves je, da ne deluje z določenimi tipi podatkov - ima probleme z zveznimi podatki, kot so npr. prihodki oz. cene. Prav tako je problem ta, da ko se algoritem odloči, da izbere eno spremenljivko kot bazno (spremenljivko, ki

je najpomembnejša), te odločitve nikoli ne spremeni. Problematični so tudi atributi brez vrednosti (ti. »NULL« vrednosti).

Nevronska indukcija je podobna nevronskim mrežam. Ponazorjena je z vozlišči in povezavami med le-temi, procesira se v vsakem vozlišču. Vsako vozlišče (procesna enota) v eni plasti je povezano z vsemi vozlišči v drugi plasti z uteženimi povezavami, ki prikazujejo moč razmerja med vozlišči. Te uteži so večinoma majhne, neničelne številke in se sproti spreminjajo, tako da nevronska mreža prilagodi izhod na pričakovani razred (na podlagi izračunov iz danih podatkov). Če se izhod razlikuje, se izračuna popravek in se ta popravek uporabi pri nadaljnjem procesiranju v vozliščih v mreži. To se odvija dokler se ne doseže končnega pogoja (ang. Stopping condition).

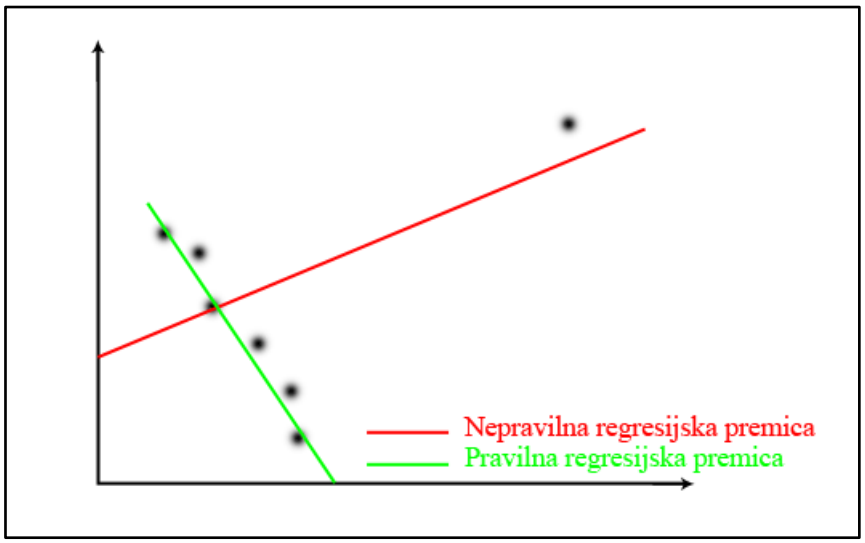


Slika 2.3. Primer nevronske mreže.

Nevronske mreže odpravljajo slabost indukcije dreves, saj vsako točko ovrednotijo tekom izračunavanja prilegajoče funkcije (ang: fitting function). Pomanjkljivost pa je predvsem ta, da nevronske mreže lahko uporabljamo predvsem pri numeričnih vhodih (za kategorične podatke moramo uporabiti katero izmed tehnik za prevedbo na numerične podatke).
Nekateri modeli nevronskih mrež ne konvergirajo k predikciji, ki jo zahteva analitik. To se dogaja predvsem zaradi nečistih podatkov in pri problemih, ki so preveč obsežni za nevronske mreže.
Nevronske mreže bi lahko rekli, da so črna skrinja [1, stran 75], saj kriteriji po katerih algoritem določi bolj pomembne dejavnike od drugih niso takoj na voljo analitiku. To lahko rešimo z dvema metodama: analizo vhodne občutljivosti (ang. Input sensitivity analysis) ali z matriko zmede (ang. Confusion matrices). S prvo metodo lahko analitik vidi, kateri so bolj pomembni dejavniki pri razdelitvi podatkov v razrede, druga metoda pa ponuja stopnjo učinkovitih klasifikacij tako, da prikaže število pravilnih in nepravilnih klasifikacij za vsako možno vrednost.

Včasih uporabnik želi raje napovedati manjkajočo vrednost kot pa razred. To je predvsem v primerih, ko so podatki numerični. Temu rečemo **predikcija.** Pri predikciji uporabljamo predvsem linearno regresijo in nelinearno regresijo. Razlika med obema tehnikama je v tem, da linearna regresija poskuša narisati ravno premico skozi med podatki na tak način, da premica kar najbolj predstavlja povprečje opazovanj.
Problem linearne regresije je, da vrača ustrezne rezultate (rezultate kakršne pričakujemo) le, če so podatki tudi sicer linearno razporejeni. Kot drugi problem pa naj izpostavim preveliko

občutljivost linearne regresije na posamezne točke, ki nekoliko odstopajo od realne premice, kar prikazuje naslednja slika.



Slika 2.4. Pomanjkljivosti linearne regresije.

Pred klasifikacijo in predikcijo je včasih smotrno prečistiti podatke, narediti analizo **ustreznosti**, ki omogoča, da identificiramo atribute, ki za proces klasifikacije oz. predikcije niso pomembni, in v skladu z zahtevami transformirati podatke.

2.3.2 Segmentacija podatkovne baze

Segmentacija podatkov je v poslovnem svetu (pri današnjih velikih količinah raznolikih podatkov v podatkovnih bazah) še posebej uporabna in pomembna pri segmentaciji strank v manjše skupine za dodatno analizo marketinških aktivnosti.
Delimo jo na demografsko razvrščanje v skupine (ang. Demographic Clustering) in nevronsko razvrščanje v skupine (ang. Neural Clustering).

Pri **demografskem razvrščanju v skupine** se množico podatkov razdeli na skupine na podlagi primerjave vsakega zapisa z obstoječimi (trenutnimi) segmenti, ki smo jih že naredili s pomočjo prejšnje iteracije algoritma za podatkovno rudarjenje. Ta tehnika se zanaša na enostavno metodo *Condorset*, ki izračuna razliko med vhodom (trenutnim zapisom) in trenutnimi skupinami, na podlagi katere dodeli zapis eni izmed skupin. Število polj, pri katerih so si zapisi podobni oz. se ujemajo, imenujemo *stopnja ujemanja*. Število polj, pri katerih se zapisi zelo razlikujejo, imenujemo *stopnja neujemanja*. Algoritem nato na podlagi števila ujemanj dodeli število točk. Glede na seštevek točk se trenutni zapis dodeli eni izmed skupin. Zapis lahko dodelimo drugi skupini, če je seštevek točk večji kot seštevek točk, če bi zapis dodelili drugi skupini. Če je seštevek točk negativen, je ta zapis kandidat za svojo skupino. V nasprotju z nevronskim razvrščanjem v skupine, ki ga bom opisal v nadaljevanju, je demografsko razvščanje v skupine primerno predvsem za kategorične podatke, kjer nastopa manjše število kategorij.

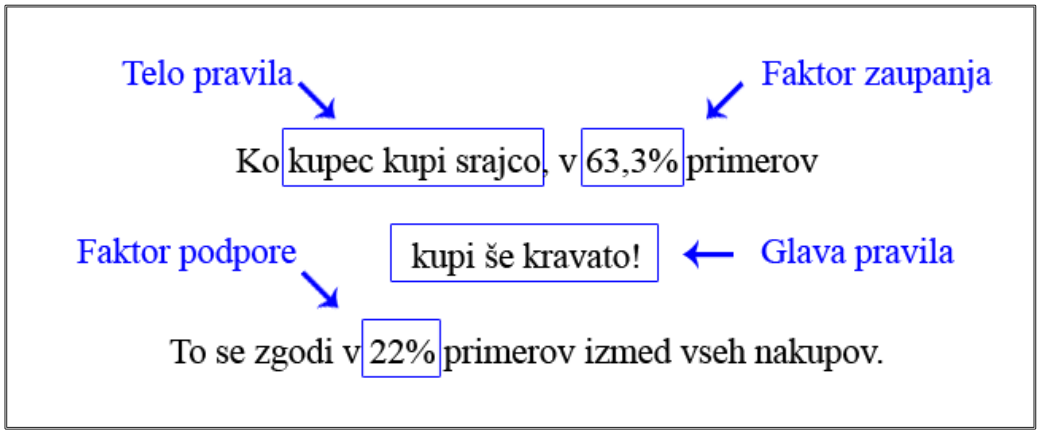
Nevronsko razvrščanje v skupine je primerno predvsem za numerične vrednosti podatkov. Uporablja nevronske mreže in uporablja značilnost načrtovanja Kohonen, ki uporablja proces samoorganizacije za nastavitev izhodnih enot v topologijo povezav (gre za nenačrtovano učenje). Nevronske mreže s to značilnostjo načrtovanja so sestavljene iz dveh plasti procesnih enot, in sicer iz vhodne plasti, ki je povezana z izhodno plastjo (je brez skritih plasti, v primerjavi z nevronskimi mrežami pri nevronski indukciji). Ko na vhod pride vzorec, izhodne enote tekmujejo druga z drugo. Zmagovalna izhodna enota je običajno enota, katere utež na vhodni povezavi je najbolj podobna vhodnemu vzorcu (običajno za primerjavo vzamemo Evklidovo dolžino). Ko je razglašen zmagovalni izhod, se mu popravi vhodna utež.
Delovanje načrtovanja Kohonen naredi topologijo povezav, katerih uteži spreminja, hkrati pa spreminja uteži na sosednjih povezavah.

2.3.3 Analiza povezav

Analiza povezav (v literaturi tudi asociacijska analiza) išče povezave med podatki, ki jih proučujemo. V to skupino algoritmov za podatkovno rudarjenje sodijo naslednje tehnike:

- Odkrivanje povezav,
- Odkrivanje zaporedja vzorcev,
- Odkrivanje podobnih časovnih zaporedij.

Namen tehnike **odkrivanja povezav** je iskanje objektov, ki implicirajo prisotnost drugih objektov v isti transakciji. Če to pretvorimo v problem realnega sveta, to pomeni, da iščemo povezave med izdelki in storitvami. Povezave med izdelki zapišemo kot asociacijsko pravilo. Analiza povezav je raziskovanje asociacijskih pravil oz. zakonitosti, ki pokažejo vrednosti atributov, ki se pogosto pojavljajo skupaj v množici podatkov (so povezani med seboj). Asociacijsko pravilo (ang. Association rule) je implikacija izraza oblike X→ Y, kjer velja X∩Y = ∅ [6]. Moč (ang. Strength) asociacijskega pravila lahko merimo z izrazi faktor podpore (ang. **Support factor**) in faktor zaupanja (ang. **Confidence factor**). Podpora določa kako pogosto je pravilo uporabljeno na množici podatkov, medtem ko zaupanje pove kako pogosto se objekti iz množice Y pojavljajo v transakcijah, ki vsebujejo X.



Slika 2.5. Asociacijsko pravilo.

$$\text{Support, } s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Slika 2.6. Definiciji podpore in zaupanja.

Podpora je pomembna mera, saj pravilo z majhno podporo lahko nastopi enostavno po naključju. Prav tako je pravilo z majhno podporo (verjetno) nezanimivo s poslovnega stališča, saj (verjetno) ne bo dobičkonosno, če bomo delali promocijo pri izdelkih, ki jih uporabniki redko kupujejo skupaj. Podporo torej uporabljamo za eliminacijo nezanimivih pravil [12].
Po drugi strani pa zaupanje meri zanesljivost dogodka, ki ga je povzročilo pravilo. Pri pravilu $X \rightarrow Y$ pomeni, da višje kot je zaupanje, bolj verjetno je, da bo Y prisoten v transakcijah poleg X. Zaupanje prav tako pomaga pri ocenitvi pogojne verjetnosti Y pri danem X.

Pri iskanju uporabnih asociacijskih pravil med transakcijami T moramo najti vsa pravila, ki imajo podporo ≥ *minsup* in zaupanje ≥ *minconf*, kjer sta *minsup* minimalna vrednost stopnje podpore in *minconf* minimalna vrednost stopnje zaupanja.

Ta tip analize je pogost pri raziskovanju košarice (ang. Market basket analysis) ali pri analizi transakcijskih podatkov transakcijskih podatkovnih baz, npr. analiziranje postavk prodaje.

Odkrivanje zaporedja vzorcev je uporabna metoda pri odkrivanju vzorcev med transakcijami, ki si sledijo v nekem obdobju. Metoda torej preučuje ali se je neka skupina izdelkov, ki se je v transakcijah pojavljala v preteklosti, pojavlja tudi v kasnejšem časovnem obdobju [1, stran 83].

Primer: Na voljo imamo podatkovno bazo za prodajo (transakcije), čas transakcije (oz. ID transakcije) in stranko (skupina transakcije). Vsaka transakcija ima skupino artiklov, ki ji pripadajo.
Iz tega lahko izpeljemo sledeče:

- Vsaka transakcija je določena s časom transakcije,
- Vsak izdelek je določen z enoličnim identifikatorjem izdelka,
- Vsaka stranka je enolično določena s številko stranke (skupina transakcije).

Stranka	Čas transakcije	Izdelki
A. Novak	21. 3. 2008 14:32	Pivo
A. Novak	22. 3. 2008 18:49	Konjak
B. Pirc	20. 3. 2008 7:09	Pomarančni sok, Kola
B. Pirc	20. 3. 2008 10:04	Pivo
B. Pirc	21. 3. 2008 12:48	Vino, Voda, Jabolčni sok
N. Sekan	21. 3. 2008 15:09	Pivo, Gin, Jabolčni sok
T. Kesovija	20. 3. 2008 14:32	Pivo
T. Kesovija	21. 3. 2008 9:08	Vino, Jabolčni sok
T. Kesovija	22. 3. 2008 17:45	Konjak
F. Zaplotnik	20. 3. 2008 11:32	Konjak

Tabela 2.1. Primer zapisov v transakcijski podatkovni bazi za prodajo.

Zgornja tabela je sortirana po stranki in po času transakcije.

Stranka	Izdelki
A. Novak	(Pivo) (Konjak)
B. Pirc	(Pomarančni sok, Kola), (Pivo), (Vino, Voda, Jabolčni sok)
N. Sekan	(Pivo, Gin, Jabolčni sok)
T. Kesovija	(Pivo), (Vino, Jabolčni sok), (Konjak)
F. Zaplotnik	(Konjak)

Tabela 2.2. Zaporedje nakupov po strankah.

Odkrivanje zaporedja vzorcev, faktor podpore > 40%	Stranke
(Pivo) (Konjak)	A. Novak, T. Kesovija
(Pivo) (Vino, Jabolčni sok)	B. Pirc, T. Kesovija

Tabela 2.3. Prikazuje izdelke, ki se pojavljajo skupaj v nakupovalnih košaricah.

Tabelo 2.3 smo dobili s pomočjo metode za odkrivanja zaporedja vzorcev in prikazuje kateri izdelki se pojavljajo v nakupovalnih košaricah (faktor podpore v zgornjem primeru je višji ali enak 40%).

Za razliko od odkrivanja povezav moramo pri metodi odkrivanja zaporedja vzorcev določiti samo en parameter, tj. faktor podpore. Izračun faktorja podpore se nekoliko razlikuje od tistega pri odkrivanju povezav, v osnovi pa sta enaka. Potrebujemo pa veliko količino podatkov o

transakcijah kupcev. Problematično pa je tudi pomanjkanje podatkov o strankah (na tak primer pogosto naletimo pri maloprodaji).

Metoda **odkrivanja podobnih časovnih zaporedij** najde vse pojavitve oz. podobne pojavitve in tudi zaporedja podobna danemu zaporedju v podatkovni bazi, kjer so podatki v obliki časovne vrste (ang. time-series data). Časovna vrsta je vrsta podatkov v posameznih časovnih trenutkih ali v zaporednih časovnih intervalih.

Dobra lastnost odkrivanja podobnih časovnih zaporedij je ta, da lahko premike več različnih poslovnih opažanj raziskujemo s podatkovnim rudarjenjem brez kakršnihkoli pogojev. Pod pojmom poslovna opažanja mislimo pojme, kot so vsota prodanih izdelkov, premiki cen izdelkov in zaloge, prodaja po lokacijah, ...

Problemi pri tej metodi pa so predvsem zahtevnost uporabe (za nove uporabnike), saj mora uporabnik nastaviti precej parametrov, npr. še dovoljeno odstopanje za primerjavo podobnih časovnih zaporedij.

> *Primer*:
> Prodajalec želi optimizirati nabavo in skladiščenje izdelkov, ki jih prodaja. Lahko analizira prodajo izdelkov oz. skupin izdelkov na dnevni ali tedenski ravni in tako lahko opazi kateri izdelki (oz. skupine izdelkov) se bolje prodajajo kot drugi (oz. druge skupine izdelkov). In nasprotno – kateri izdelki se slabše prodajajo kot drugi izdelki. V njegovi podatkovni bazi je prodaja zapisana kot vrsta zaporednih postavk, ki predstavljajo zmanjševanje in povečevanje (ob nabavi, reklamacijah) količine izdelkov v časovnem intervalu.
> To funkcionalnost zajema že prej omenjena aplikacija za poslovno obveščanje. Algoritem za podatkovno rudarjenje, ki ga poženemo nad temi podatki, pa nam prikaže vsa podobna zaporedja gibanja izdelkov. S pravilno interpretacijo teh informacij lahko prodajalec najde skupine izdelkov, ki imajo podobna sezonska gibanja prodaje za sledeča leta. Tako lahko tudi bolje napove nabavo izdelkov in dopolnitve zalog.

2.3.4 Odkrivanje odstopanj

Odkrivanje odstopanj išče razlike oz. odstopanja v podatkih, ki se razlikujejo od pričakovanj. Pod skupino odkrivanje odstopanj sodita metoda **vizualizacije** in **različni statistični pristopi**. Pri metodi vizualizacije gre predvsem za prikaz informacij, ki smo jih pridobili iz podatkov s pomočjo podatkovnega rudarjenja (prikažemo rezultat podatkovnega rudarjenja).

Metoda **vizualizacije** je ena izmed bolj uporabnih tehnik odkrivanja vzorcev v podatkih [13, 14]. To tudi ni nič čudnega, saj psihologi pravijo, da je 80% informacij, ki jih človek sprejme, sprejetih s pomočjo vida [1, stran 86]. Vizualizacija je še posebno koristna pri iskanju pojavov, ki vsebujejo majhne podmnožice podatkov in jih običajno ne opazimo zaradi prevlade drugih močnejših podmnožic podatkov, ki jih proučujemo.

Poznamo več načinov vizualizacije:

- za predstavitev podatkov glede na eno spremenljivko oz. dimenzijo so uporabni predvsem naslednji načini predstavitev:
    - histogrami,
    - raztreseni diagrami (ang. scatterplot),
    - škatlasti diagrami (ang. boxplot),
    - krožni grafikon (ang. pie chart).
- za eno, dve ali tri spremenljivke oz. dimenzije so uporabni predvsem tridimenzionalni grafikoni.
- za večdimenzionalne probleme pa nastajajo vsak dan novejši grafikoni.

Dobra lastnost vizualizacije je ta, da nam ni potrebno vnaprej določiti domnev, da bi s pomočjo vizualizacije odkrili nekaj zanimivega ali neobičajnega.

Vizualizacija je predvsem dopolnilo ostalim tehnikam za podatkovno rudarjenje, saj lahko z njeno pomočjo na enostaven in hkrati zelo zgovoren način uporabniku predstavimo rezultate podatkovnega rudarjenja.

Poleg vizualizacije poznamo pri odkrivanju odstopanj še **statistične pristope**. Ti pristopi so bolj teoretski in so bolj namenjeni za testiranje hipotez, zato jih ne bom toliko opisoval.

2.4 Pomembne tabele, nad katerimi bomo izvedli podatkovno rudarjenje

Za izvedbo podatkovnega rudarjenja se bomo povezali na Microsoft SQL strežnik, kamor Microsoft Dynamics NAV shranjuje podatke o transakcijah v poslovnem okolju. Iz ERP sistema Microsoft Dynamics NAV bom prenašal podatke o prodaji, ki so shranjene v spodaj opisanih tabelah dejstev.

Tabele dejstev:

Št. postavke	Datum knjiženja	Tip postavke	Št. artikla	Količina	Šifra lokacije	Preostala količina
13	1. 3. 2008	Nakup	2930002	5	01	1
378	3. 3. 2008	Prodaja	2930002	-2	01	0
501	19. 3. 2008	Prodaja	2930002	-2	01	0

Tabela 2.4. Primer postavk za artikel 2930002 iz tabele Postavka artikla.

Št. postavke	Št. postavke artikla	Znesek stroška (dejanski)	Datum knjiženja
22	13	12,45	1. 3. 2008
51	13	-0,4	5. 3. 2008

Tabela 2.5. Primer postavk iz tabele Postavka vrednosti za izbrane postavke iz Postavk artikla.

Št. postavke	Št. postavke artikla	Znesek stroška (dejanski)	Datum knjiženja
22	13	12,45	1. 3. 2008
51	13	-0,4	5. 3. 2008

Tabela 2.6. Primer postavk iz tabele Postavka kupca.

Št. postavke	Št. postavke artikla	Znesek stroška (dejanski)	Datum knjiženja
22	13	12,45	1. 3. 2008
51	13	-0,4	5. 3. 2008

Tabela 2.7. Primer postavk iz tabele Podrobna postavka kupca.

Razširjene tabele se nahajajo v Prilogi.

Šifranti:

Prikazane tabele niso standardne Microsoft Dynamics NAV tabele, ampak sem jih razširil s podatki, ki so zanimivi za izvedbo podatkovnega rudarjenja za namene tega diplomskega dela.

V povezavi z obstoječo aplikacijo za poslovno obveščanje za poslovno funkcijo nabave bi bilo uporabno povezati se na bazo AJPES [15], kjer se nahajajo pravni subjekti v Republiki Sloveniji. Od tam bi bilo uporabno izvoziti različne podatke in jih dodati bodisi dodati v svoj podatkovni model bodisi uvoziti v podatkovno skladišče preko ETL procedur.
Za potrebe tega diplomskega dela sem uvozil sledeče podatke:

- velikost podjetja (mikro*, malo**, srednje veliko***, veliko**** [9]); če je končni kupec, piše »končni«,
- čas obstoja podjetja,
- datum ustanovitve podjetja,
- osnovni kapital,
- število zaposlenih.

Poleg zgoraj navedenih podatkov je dobro hraniti tudi podatke, ki nastajajo sčasoma, ko podjetje sodeluje z zunanjimi podjetji, npr:

- Kdaj je podjetje postalo stranka?
- Zamude pri plačilih
- Povprečen čas plačil

Z združitvijo v en sistem bi lahko iz teh podatkov s pomočjo podatkovnega rudarjenja izvedeli veliko zanimivih in uporabnih informacij, kot so:

- Ali obstajajo zakonitosti pri plačevanju računov s strani mikro, majhnih, velikih podjetij in kakšne so te zakonitosti.
- Ali podjetja, s katerimi sodelujemo že dolgo bolje plačujejo kot tista, ki so pravkar postala stranka?
- Ali mlada podjetja bolje plačujejo kot starejša?
- Kakšna je disciplina naročanja glede na zgornje parametre?

Šifra kupca	Naziv	Naslov	ZIP	Kraj	Velikost podjetja	Čas obstoja	Osnovni kapital
K1082126112	Kupec K1082126112	TOVARNIŠKA 12	5270	AJDOVŠČINA	mikro	2	8500
K1082126878	Kupec K1082126878	KALE 184	3320	VELENJE	veliko	71	112500
K1016125510	Kupec K1016125510	SLOVENSKA 55	6400	LJUBLJANA	mikro	13	12000
K1016125530	Kupec K1016125530	Stegne 19	1400	LJUBLJANA	končni	53	-

Tabela 2.8. Primer postavk iz tabele Kupec.

Šifra artikla	Naziv	Knjižna skupina artikla	Strošek enote	Posreden strošek	Zadnji neposredni strošek
2930002	Svinčnik	115300	0,53	0,02	0,52
1152201	Nalivno pero	115300	12,33	0,13	12,31
1355549	Radirka	115300	0,13	0,01	0,13

Tabela 2.9. Primer postavk iz tabele Arikel.

Številka	Ime in priimek	Delovno mesto	Naslov	Kraj	ZIP
1	Delavec 1	SKLADIŠČNIK	PREGLOV TRG 2	LJUBLJANA	1000
2	Delavec 2	VODJA ODD.FINANC	BEBLERJEV TRG 6	LJUBLJANA	1000
10	Delavec 3	PRODAJNI REFERENT	PREŠERNOVA 33	LJUBLJANA	1000
11	Delavec 4	PRODAJNI REFERENT	BUKOVICA 43	VODICE	1217
12	Delavec 5	PRODAJNI REFERENT	DRAŽGOŠKA 7	KRANJ	4000
23	Delavec 6	PRODAJNI REFERENT	CELOVŠKA 189	LJUBLJANA	1000

Tabela 2.10. Primer postavk iz tabele Delavec.

Razširjene tabele se nahajajo v Prilogi.

3 Podatkovno rudarjenje kot način za povečanje dodane vrednosti

Podatkovno rudarjenje lahko najbolje uporabimo predvsem na področju marketinga in načrtovanja prodaje. Spodnja tabela nekoliko bolj podrobno razčlenjuje področja uporabe operacij in tehnik podatkovnega rudarjenja.

Uporabnost posameznega algoritma je pogojen z naravo problema, ki ga želimo dodatno raziskati.

Področje	Marketing	Upravljanje s tveganjem	Nadzor nad prevarami
Aplikacije	Direktni marketing Ciljani marketing CRM Analiza nakupovalne košarice Navzkrižna prodaja Segmentacija trga	Napovedovanje Ohranitev kupca Izboljšano preverjanje sposobnosti odplačevanja stranke Kontrola kvalitete Analiza tekmecev	Detekcija prevar

Operacije	Napovedno modeliranje	Segmentacija podatkov	Analiza povezav	Odkrivanje odstopanj
Tehnike	Klasifikacija Napovedovanje vrednosti	Demografsko razvrščanje Nevronsko razvrščanje	Odkrivanje povezav Odkrivanje zaporedja vzorcev Odkrivanje podobnih časovnih zaporedij	Vizualizacija Statistika

Tabela 3.1. Področja uporabe podatkovnega rudarjenje in operacije ter tehnike, ki ta področja podpirajo.

3.1 Uporaba različnih algoritmov za podatkovno rudarjenje po različnih področjih v prodaji

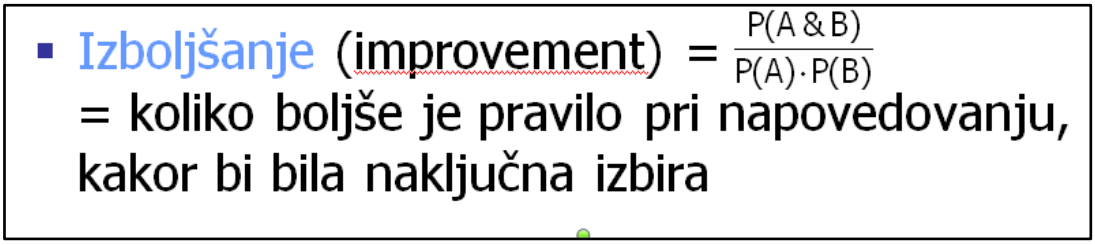
Vsebinsko lahko področja za izvedbo podatkovnega rudarjenja razdelimo na naslednje sklope:
- Marketing (ang. Marketing management),
- Upravljanje s tveganjem (ang. Risk management),
- Nadzor nad prevarami (ang. Fraud management).

3.1.1 Marketing

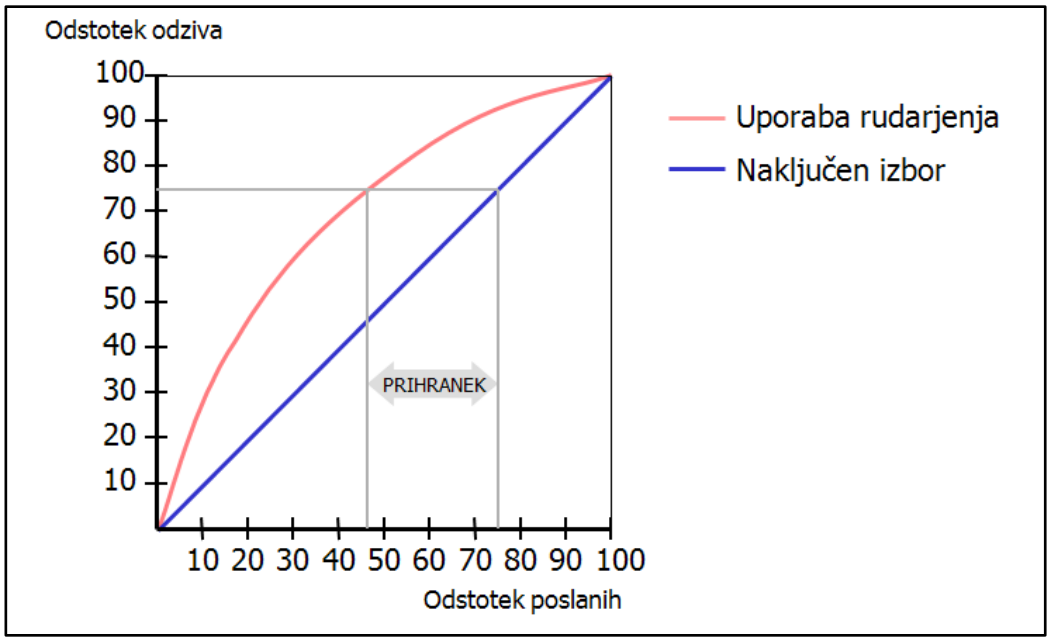
V sklopu marketinga lahko na prodajo vplivamo z naslednjimi prijemi:
a) Direktni marketing,
b) Ciljani marketing,
c) Vpeljava in povezovanje na CRM sistem,
d) Analiza nakupovalne košarice,
e) Navzkrižna prodaja (ang. cross-sell),
f) Segmentacija trga.

Pri **direktnem marketingu** gre za proces neposrednega kontaktiranja stranke (npr. pošiljanja pošte, klicanje, osebni obisk strank, ki so v to privolile). Tudi tu lahko stranke, ki so privolile v prejem reklamnih materialov, razdelimo v segmente (v večini primerov pa se že same pri prijavi odločijo za tematiko materialov, ki jo želijo prejemati).

Izboljšava direktnega marketinga
Veliko podjetij še vedno dela veliko napako, da s pomočjo direktnega marketinga ciljajo na celotno populacijo. To pomeni, da pošiljajo (neželeno) pošto na vse naslove, ki jih imajo v bazi ali pa gredo še korak naprej – dajejo reklame v vsak nabiralnik. Znano je, da je pozitivnih odgovorov na tak marketing manj kot 1% [7], kar pomeni, da podjetje, ki se poslužuje takšnega načina marketinga, veliko denarja porabi na nesmotern način.
Kako lahko že s sorazmeno majhnim vložkom v IT izboljšamo ROI?
Z izboljšavo direktnega marketinga lahko izboljšamo odgovore (potencialnih) strank na prodajne akcije, ki jih izvajamo trenutno. Vsi marketinški specialisti namreč dajejo velik poudarek na dobro organizirano podatkovno bazo. Pravijo, da je v podatkovni bazi potrebno imeti kvalitetne podatke o strankah.



Slika 3.1. Slika prikazuje izboljšanje pri napovedovanju zaradi uporabe podatkovnega rudarjenja.



Slika 3.2. Prihranek z uporabo podatkovnega rudarjenja.

Pri **ciljanem trženju** gre za proces marketinga, kjer pošiljamo reklamni material natanko določenemu segmentu kupcev (ciljamo končno skupino ljudi). Uspeh prodajne akcije, ki je osnovana na ciljanem trženju, je zelo odvisna od dobrega algoritma, ki naredi pravilne segmente strank. Pri ciljanem trženju je zelo pomembno komu pošljemo reklamni material oz. ponudbo, saj se večji del ljudi na ponudbe ne odzove. S tem ko skrbno izberemo množico, kateri pošiljamo ponudbe, vplivamo tako na zmanjšanje stroškov (zaradi tiska reklamnih materialov, poštnine, ...) kot tudi na povečanje odziva s strani naslovnikov (v večini primerov je ciljana množica skrbno izbrana množica ljudi, ki so potencialni kupci naših izdelkov ali storitev).
Pri prodaji se velikokrat naslanjamo na podatke o kupcih, ki jih dobimo v sistemu **CRM**. Podjetja morajo kot del svoje poslovne strategije razviti portfeljsko obliko trajnejših razmerij s kupcem, pri čemer poglobljenost razmerij določata obseg poslovanja s kupcem in vsebina (poslovni učinek, razpečava in logistika, oglaševanje, cena). Kakovosten CRM sistem vodi k povečevanju tržnega deleža na sedanjih trgih, osvajanju novih trgov, krajšim prodajnim lokom ter nižjim stroškom prodaje. Kakovosten CRM podjetju omogoča dobro poznavanje stranke in boljše ciljanje pri prodajnih akcijah. Zniževanje stroškov pa nadalje vodi k bolj učinkovitemu ter bolj dobičkonosnemu poslovanju.
Podatkovno rudarjenje je še posebej učinkovito v povezavi s CRM sistemom, saj lahko s povezavo teh dveh konceptov bolje spoznamo naše potrošnike in trg, kar posledično pripelje do večje učinkovitosti prodaje.

Analiza nakupovalne košarice je v časih, ko se Internet čedalje bolj razvija, ključnega pomena. Namenjena je predvsem pri elektronskem poslovanju (ang. e-commerce), saj v tem primeru pogosto imamo podatke o strankah in vemo kaj so kupile.
Predvsem e-prodajo (ang. e-commerce) pa tudi klasično prodajo lahko nadgradimo tako, da stranki ponudimo artikle, ki jih stranke pogosto kupujejo sočasno. To pomeni, da naredimo analizo nad dosedanjimi nakupi strank in jim na ta način ponudimo še kakšen podoben artikel ali pa ji ponudimo boljšo različico artikla, ki ga ima namen kupiti (ang. up-sell).

Navzkrižna prodaja je zadnje čase zelo pogosta zahteva pri implementaciji programske opreme za prodajo. Navzkrižna prodaja pomeni, da stranki, ki kupi določen izdelek ponudimo še izdelek oz. storitev, ki bi jo verjetno želela kupiti [16]. To je z razvojem algoritmov za podatkovno rudarjenje doživelo svoj razcvet, saj ti algoritmi omogočajo (enostavno) iskanje podobnih transakcij v podatkovnih bazah (uporabimo lahko različne tehnike analize povezav).

Segmentacija trga je zelo uporaben prijem, saj omogoča podjetju, da lahko razdeli svoje stranke na segmente. Za posamezne segmente lahko pripravlja različne ponudbe, ki stojijo na predpostavkah, da so si znotraj segmentov stranke podobne in imajo podobne želje oz. zahteve. Segmentacija trga je pogosto pomemben člen pri izdelavi marketinških kampanj, saj je podlaga za direktni in ciljani marketing in tudi za razumevanje strankinih navad in običajev.

V zadnjem času pa se podatkovno rudarjenje vse bolj uporablja tudi za **spletno rudarjenje** (ang. web mining), ki uporablja tehnike podatkovnega rudarjenja za odkrivanje vzorcev na Internetu. Delimo ga na tri sklope:

- **Rudarjenje po uporabi spletnih datotek** (ang. web usage mining), ki se osredotoča na iskanje vzorcev uporabe spletnih vsebin. Če shranjujemo podatke o obiskanih spletnih straneh posameznega uporabnika (z vsakim klikom uporabnik sproži transakcijo proti strežniku), lahko analiziramo in iščemo vzorce v obnašanju uporabnika – to lahko pripelje podjetje do vrednih podatkov o določenih strankah, lahko izdelamo strategije navzkrižne prodaje in bolj efektivno izpeljemo promocijske kampanje.
- **Rudarjenje po vsebini na spletu** (ang. web content mining) je proces odkrivanja uporabnih informacij iz tekstov, slik, avdio in video datotek, ki so dostopne preko svetovnega spleta. Največkrat gre za raziskovanje teksta, zato mu pravimo tudi **web text mining**. Najpogosteje uporabljane tehnike so procesiranje naravnega jezika (NLP) in pridobivanje informacij (IR).
- **Rudarjenje po strukturah na spletu** (ang. web structure mining) je proces analiziranja struktur. Uporablja teorijo grafov za analizo vozlišč in povezav in na ta način ugotovi strukturo spletne strani. Poznamo 2 tipa strukturnega rudarjenja. Prvi tip deluje na principu ekstrahiranja vzorcev iz hiperpovezav na spletni strani. Drugi pa je tak, da pregleda strukturo dokumenta. Ta tip uporablja drevesno strukturo za analizo in opis HTML oz. XML značk znotraj strani.

3.1.2 Upravljanje s tveganjem

Pri upravljanju s tveganjem skuša podjetje zmanjšati tveganje zaradi različnih možnih vplivov okolja na poslovanje. Pod upravljanje s tveganjem sodijo napovedovanje (prodaje), ohranitev kupca, izboljšanje preverjanja sposobnosti odplačevanja stranke, kontrola kvalitete in analiza tekmecev. Več znanja kot imamo o teh pojmih, manjše tveganje nosi podjetje.

Napovedovanje
S pomočjo napovednega modeliranja lahko dobro napovedujemo gibanje prihodkov in odhodkov, dobička, prodaje (količinsko) v prihodnosti. To je tudi podlaga za boljše planiranje (ang. budgeting), ki je danes zelo pomemben člen poslovnega sveta. Vsa večja podjetja namreč

planirajo svojo stroške, prodajo in nabavo, saj jim to velevajo njihovi lastniki. Te plane potem primerjajo z realiziranimi, kar je dober kazalnik o uspešnosti podjetja. Če podjetje dosega oz. presega te plane, je to dobro in obratno če jih ne dosega, je slabo.
Pri napovedovanju je vse bolj pomembno tudi napovedovanje stanja zalog. Zaloge so zelo kompleksen pojem. Stremeti moramo k temu, da imamo ravno pravšnjo zalogo izdelkov v skladišču, ne preveč ne premalo. Preveč izdelkov pomeni slabost, ker imamo manj denarja, ki ga lahko obračamo, premalo izdelkov pa pomeni, da ne moremo pravočasno izpolniti naročila. S pomočjo podatkovnega rudarjenja lahko bolj natančno napovemo (glede na pretekle trende – prodajo izdelka v prejšnjih letih) kakšna bo povpraševanje in lahko pravočasno izdelke dodamo na zalogo, v kolikor nam jih primanjkuje.

Ohranitev kupca
V kolikor imamo na voljo podatke o zadnjih nakupih naših strank, lahko napovemo, kako se bo stranka obnašala v prihodnje. Tu je dobro, da imamo v CRM sistemu čim več podatkov – vključno z demografskimi, ki omogočajo natančnejše in boljše analize strank (segmentacijo). Na podlagi obstoječih podatkov iz podatkovnega skladišča lahko napovemo, za katere strank je še posebej visoka verjetnost, da bodo »prebegnile« (odšle) h konkurenci zaradi visokih cen, slabih plačilnih pogojev, kvalitete izdelkov, ... Stranko moramo tudi vprašati, zakaj je odšla drugam in to zapisati v CRM, saj lahko le s (približno) popolnimi podatki izdelamo boljše profile strank v odhajanju.
Ocenimo lahko **vrednost stranke** in kakšna je pravilna ponudba, da bo stranka določen izdelek kupila pri nas. Pravilno ponudbo pa je pogosto težko napovedati, zato se današnji prodajni referenti še vedno nagibajo k intuiciji pri sestavljanju ponudb za stranke. S pomočjo podatkovnega rudarjenja lahko na podlagi dejstev določimo vrednost ponudbe. S tem, ko bolje napovemo vrednost, ki bi jo bila stranka še pripravljena plačati za izdelek pri nas, lahko bolje določimo ceno in ji posledično tudi več prodamo. Več kot prodamo, večji je dobiček, le-ta pa je gonilo današnjih kapitalističnih družb.

Bela knjiga: Verizon Wireless
Verizon Wireless je leta 2003 zgradil podatkovno skladišče za stranke. Podatkovno skladišče je nastalo z namenom preprečitve prebega strank. Razvil je več regionalnih modelov. Na podlagi segmentacije strank je natančno določil segmente strank. Te segmente je lahko na podlagi podatkov v podatkovnem skladišču natančno določil in jih »naciljal« s ponudbo, ki so jo z veliko verjetnostjo sprejeli [13, 17].
Veriozon Wireless je največji brezžični operater v ZDA, ki je imel bazo naročnikov s 30,3 milijoni naročniki. S svojimi storitvami je pokrival 90% prebivalstva. Izračun kaže, da jih pridobitev nove stranke stane 320 dolarjev. Za nadomestitev izgube zaradi prebega strank so tako na leto izgubili več sto milijonov dolarjev.

S pomočjo napovednega modeliranja so za vsako stranko določili možnost prebega in možnost, da bi stranka reagirala na ponudbo, ki so ji jo poslali. Rezultat so uporabili za ciljanje strank s specifičnimi, relevantnimi in časovnimi ponudbami.

Pri Verizon Wireless so s pomočjo podatkovnega rudarjenja naredili sledeče:

- stranki so ponudili novo spodbudo v obliki akcije zato, da je podaljšala obstoječe naročniško razmerje,
- kontaktirali so stranko, da bi s tem zmanjšali možnost prebega.

Pri Verizon Wireless so se naučili, da je pomembno, da IT in marketing oddelek delata skupaj. V njihovem primeru je IT oddelek prišel do marketinga in jim predstavil idejo kot partnerstvo. Ljudje iz marketinga so se naučili proces modeliranja hkrati pa prednosti in slabosti modeliranja, IT pa se je naučil poslovnih procesov in strategij direktnega marketinga. Marketing je nato predlagal dodatne parametre za predikcijo pri izgradnji modela.
S pomočjo podatkovnega rudarjenja je Verizon Wireless zmanjšal prebeg strank z 2%/mesec na 1,5%/mesec, kar je velik dosežek pri 30 milijonov strank. S tem so pri Verizon Wireless privarčevali več sto milijonov dolarjev [17]. Iz tega primera je razvidno, da je strošek ohranitve stranke vedno nižji od stroška pridobitve nove stranke.

Ocene o obnašanju kupcev lahko podamo le z določeno verjetnostjo, kar pomeni, da vsebuje vsak načrt prodaje neko določeno raven tveganja. To raven tveganja lahko precej uspešno zmanjšujemo s podatkovnim rudarjenjem.

Izboljšano preverjanje sposobnosti odplačevanja stranke
Na podlagi podatkov o prodaji in plačilih lahko pri prodaji na kredit (prodaja, pri kateri stranka dobi blago vnaprej, plača šele v nekem določenem času, ti. datum zapadlosti je tisti, ki določa, kdaj mora biti račun plačan) bolje ocenimo ali je stranka dober plačnik ali ne. Tako bi lahko razdelili stranke v skupine, ki plačujejo v roku, plačajo pred pretekom roka, plačajo z določeno zamudo, itd. Z razčlenitvijo strank v skupine bi lahko novo stranko že takoj uvrstili (na podlagi profila stranke) v razred, ki ji najbolj ustreza. To je seveda treba tehtno premisliti kateri parametri so ključni pri razvrstitvi v določen razred. Napak bi bilo namreč razvrstiti stranko v krog slabih plačnikov, ker je imela druga stranka (ki je slab plačnik) tako kot trenutna stranka poštno številko 1350. Ne moremo namreč reči, da so vse stranke s poštno številko 1350 slabi plačniki. Ta problem se s številom strank zmanjšuje in pomen podatkovnega rudarjenja se veča.
Če je stranka klasificirana kot dober plačnik (vse plačuje v rokih oz. ne prekorači datuma zapadlosti), mu lahko lažje in hitreje odobrimo naročilo in mu ponudimo boljše plačilne pogoje in morda tudi višje popuste.
Na podlagi dobrih podatkov lahko zgradimo tudi modele strank s pomočjo različnih algoritmov za strojno učenje.

Kontrola kvalitete je pomemben člen v proizvodnji izdelkov. Rezultat kontrole je lahko ena izmed konkurenčnih prednosti na vse hitreje rastočih trgih. Stranke pogosto ne kupujejo več najcenejših izdelkov, ampak se jih čedalje več odloča za kvaliteto. Kvaliteto izdelka lahko merimo tako, da zapisujemo in hranimo podatke o reklamacijah, ki jih podajo stranke. Na podlagi teh podatkov lahko ugotovimo, kateri izdelki so bolj kakovostni kot drugi. V primeru manj kakovostnih izdelkov se nato lahko odločimo ali bomo izdelek izboljšali ali jih bomo raje dali v izdelavo k drugemu proizvajalcu. Izdelke, ki bi jih drugi izvajalci naredili bolje (in morebiti tudi ceneje), je bolje kupovati pri drugemu izdelovalcu.
Seveda pa je kontrola kvalitete kompleksna stvar in presega meje tega diplomskega dela.

Analiza tekmecev je v današnjem času zelo pomembna. Rezultat, ki ga dobimo z analizo tekmecev, lahko prinese konkurenčno prednost, če jih vključimo v poslovni proces in upoštevamo pri delovanju, razvoju, cenovni politiki, itd. V današnjem svetu je potrebno dobro poznavanje svojih tekmecev. Podatkovno rudarjenje lahko pri analizi tekmecev uporabimo predvsem tako, da si podjetje zgradi podatkovno skladišče, v katerega shranjuje podatke o tekmecih. Podatki o tekmecih sestojijo iz podatkov o njihovih izdelkih in storitvah, cenah le-teh, podatke o konkurenčnih podjetjih, njihove prednosti in slabosti. Na podlagi teh podatkov lahko izluščimo uporabne informacije, ki nam dajo informacijo o poziciji podjetja glede na tekmece. Tu lahko uporabimo tudi spletno rudarjenje, kar nam omogoča, da algoritmi za podatkovno rudarjenje enostavno obiščejo spletne strani tekmecev in iz njih prenesejo želene vsebine v naše podatkovno skladišče.

3.1.3 Nadzor nad prevarami

Detekcija prevar (ang. fraud detection) je zlasti popularna v zadnjem času, ko količina kriminala narašča. S pomočjo detekcije prevar lahko odkrivamo prevare, ki se dogajajo v poslovnem okolju. Detekcija prevar za podjetja, ki se ukvarjajo s prodajo izdelkov in storitev, je zelo pomembna, še zlasti za podjetja, ki se ukvarjajo s prodajo preko Interneta. Število prevar s kreditnimi karticami se namreč s povečevanjem nakupovanja preko Interneta iz leta v leto povečuje.

Verjetno bi lahko našli prevare tudi pri prodaji končnemu kupcu, naj omenim dva primera:

- Podjetje, ki ima svojo kartico zvestobe, bi lahko spremljalo ali obstaja stranka, ki kupuje in stalno vrača oz. zamenjuje izdelke oz. se stalno pritožuje. Pri poslovanju s takšno stranko je dobro, da se podjetje vpraša, ali je smotrno poslovati z njo še naprej ali pa gre morda samo za zaporedje prevar s strani stranke, ko stranka želi npr. vedno znova menjavati izdelke za novejše.
- Podjetje zaposluje človeka, ki vedno (pri vsaki transakciji) stranki prizna popust. Če gre tu za majhne popuste, se običajno delodajalec ne bi toliko pritoževal. Če pa gre za večje popuste, pa to že zmanjšuje prodajo ter tako vpliva na poslovanje podjetja – zmanjšuje namreč dobiček podjetja. Podjetje v tem primeru vrši kontrolo tudi nad svojimi zaposlenimi in lahko hkrati ocenjuje tudi uspešnost posameznega prodajalca.

Z detekcijo teh prevar si prihranimo veliko časa in stroškov, saj nam, če odkrijemo da gre za prevaro, na primer ni potrebno pošiljati izdelkov in tako ne plačamo niti pošiljanja. Lastnik kartice se bo zelo verjetno pritožil in tako bo veliko časa in stroškov nastalo tudi zaradi birokracije.
Poleg prevar s kreditnimi karticami poznamo tudi prevare pri telekomunikacijskih podjetjih, pri preprečevanju pranja denarja, pri trgovanju z delnicami, pri iskanju terorističnih mrež...

3.2 Vertikalne rešitve

Na ERP sisteme je treba gledati kot na osnovne poslovne sisteme, ki zadovoljujejo vse osnovne poslovne funkcije podjetja. Tak sistem pa skoraj nikoli ne vsebuje modulov s katerimi bi bilo možno pokriti specifične poslovne ali tehnološke procese.
Vertikalne rešitve so dodatne rešitve, ki so integrirane z obstoječim ERP sistemom in zadovoljujejo potrebe specifične dejavnosti. Običajno jih razvijajo podjetja, ki dobro poznajo delovanje posamezne panoge ali specifičnega področja poslovanja.
Poleg naštetih zgornjih treh sklopov podatkov se v zadnjem času pojavlja vedno več vertikalnih rešitev za podatkovno rudarjenje:

- rešitve za industrijo,
- rešitve za bančništvo,
- rešitve za telekomunikacijska podjetja,
- rešitve za maloprodajo,
- ...

Te vertikalne rešitve so večinoma različne kombinacije zgoraj naštetih tehnik za podatkovno rudarjenje, ki se pogosto uporabljajo v panogi. Poleg obstoječih tehnik morajo pogosto podjetja, ki razvijajo aplikacije za podatkovno rudarjenje, razviti še dodatne tehnike in poglede na podatke za vsako podjetje posebej.

4 Podatkovno rudarjenje v orodju Microsoft Analysis Services

Orodje Microsoft Analysis Services je v paketu Microsoft SQL Server 2005, ki uporabnikom omogoča [18]:

- **pisanje MDX skript**, ki omogočajo nov mehanizem za definiranje izračunanih mer,
- **čarovnike za izdelavo aplikacij za poslovno obveščanje**,
- **enostavno izdelavo sistema uravnoteženih kazalnikov** (ang: balanced scorecard) s pomočjo KPI ogrodja,
- **podatkovno rudarjenje**, ki omogoča iskanje vzorcev in pravil v podatkih, kar omogoča predikcijo in razumevanje stvari, ki se dogajajo v poslovnem sistemu.

Podatkovno rudarjenje bom izvajal z orodje Microsoft Analysis Services zaradi več razlogov:

- večina uporabnikov trenutne aplikacije za poslovno obveščanje že uporablja Microsoft SQL Server, se pravi lahko Microsoft Analysis Services uporabljajo zastonj (zastonj verzija ima omejitev količine podatkov),
- ima uporabniku prijazen uporabniški vmesnik,
- omogoča priklop neposredno na podatke v podatkovno bazo na Microsoft SQL Server, kar pomeni, da za analizo ne potrebujemo podatkovnega skladišča.

Ker je tema tega diplomskega dela podatkovno rudarjenje, se bom osredotočil na opis funkcionalnosti in prikaz delovanja implementiranih tehnik podatkovnega rudarjenje, ki jih omogoča orodje Microsoft Analysis Services.

Tehnike podatkovnega rudarjenja v Microsoft Analysis Services
Microsoft Association Rules
Microsoft Clustering
Microsoft Decision Trees
Microsoft Linear Regression
Microsoft Logistic Regression
Microsoft Naive Bayes
Microsoft Neural Network
Microsoft Sequence Clustering
Microsoft Time Series

Tabela 4.1. Tehnike podatkovnega rudarjenja, ki jih omogoča Microsoft Analysis Services.

Opis tehnik podatkovnega rudarjenja se nahajajo v poglavju *Tehnike podatkovnega rudarjenja*. Ime Microsoft v seznamu zgornjih tehnik pomeni samo, da v ozadju teče algoritem, ki ga je napisalo podjetje Microsoft in je optimiziran za izvajanje na Microsoft SQL Server 2005.

V tem diplomskem delu se bom osredotočil predvsem na obravnavo marketinga. To pa iz preprostega razloga: podatki, ki jih v standardnem Microsoft Dynamics NAV hranijo podjetja, niso primerni oz. na voljo za izvedbo podatkovnega rudarjenja za detekcijo prevar in nekaterih ostalih potencialno zanimivih raziskav (merjenje kvalitete izdelkov, analiza nakupovalne košarice, web mining, analiza tekmecev).

Če bralca zanima, kako se dela z orodjem Microsoft Analysis Services, priporočam ogled priročnika na Microsoft spletni strani [19].

4.1.1 Navzkrižna prodaja, analiza nakupovalne košarice

Uporabili bomo naslednje tehnike:
- Microsoft Association Rules.

Podatki se nahajajo v tabelah:
- Postavka kupca,
- Podrobna postavka kupca.

Trajanje podatkovnega rudarjenja:
- Osvežitev podatkov: 8min 42s

4.1.1.1 Pridobljeno znanje

Rezultat, ki ga pridobimo z zgoraj navedenimi tehnikami (na naših nekoliko spremenjenih podatkih), je zapisan spodaj.

Support	S.	Itemset
868	2	12500 = Existing, 11900 = Existing
446	2	12200 = Existing, 12100 = Existing
4	2	12800 = Existing, 12200 = Existing
3	2	11171 = Existing, 12200 = Existing
2	2	11115 = Existing, 12200 = Existing
1	2	19000 = Existing, 12200 = Existing
868	1	11900 = Existing
868	1	12500 = Existing
482	1	12100 = Existing
456	1	12200 = Existing
438	1	19000 = Existing
16	1	12800 = Existing
12	1	11171 = Existing
8	1	11115 = Existing

Slika 4.1. Izdelki, ki nastopajo skupaj pri nakupih.

Komentar k sliki:

Videli smo, da izdelka 12500 in 11900 skupaj nastopata v 868 računih, kar pomeni, da jih kupci pogosto kupujejo skupaj v 19,4% primerov. Za izdelka 12200 in 12100 velja, da na računu nastopata v 446 primerih v 10% primerov.
Zgornji podatki so narejeni na postavkah računov, kjer stranke večinoma kupujejo samo en produkt naenkrat. Prav zaradi tega razloga tudi pride do nekoliko manj zanimivih rezultatov. Nam pa zgornja slika lahko pomaga pri postavitvi izdelkov v trgovini. V zgornjem primeru bi bilo pametno postaviti skupaj izdelke 12500 in 11900, 12200 in 12100, 12800 in 12200, 11171 in 12200, 11115 in 12200, 19000 in 12200.
Iz zgornjih podatkov je razvidno tudi, da je najbolj prodajani artikel artikel s številko 12200.
Spodnja slika prikazuje smiselno postavitev izdelkov v trgovini.

4.1.1.2 Uporabnost Microsoft Association Rules algoritma

Za uporabo Microsoft Association Rules algoritma je pomembno, da imamo kakovostne podatke. Uporaba algoritma ni smiselna, če prodajamo (večinoma) samo en artikel (račun ima samo eno postavko).

S pomočjo zgornjega algoritma lahko predvidimo tudi pametnejšo postavitev izdelkov v trgovini. V primeru prodaje preko Interneta lahko uporabniku, ki izbere določen izdelek prikažemo še izdelke, ki jih zraven pogosto kupijo ostale stranke.

Pametnejšo postavitev izdelkov v trgovini raziskuje upravljanje kategorij (ang. Category Management). Upravljanje kategorij uporabljajo predvsem različni supermarketi za izboljšano uporabnost polic v svojih poslovalnicah. Seveda pa bi lahko to razširili tudi na ostale manjše trgovine. Potrebno pa je tudi omeniti, da je postavitev različna pri različnih vrstah dejavnosti, tako je odvisno ali prodajamo prehrambene izdelke ali prodajamo drage tehnične izdelke.

4.1.2 Segmentacija trga

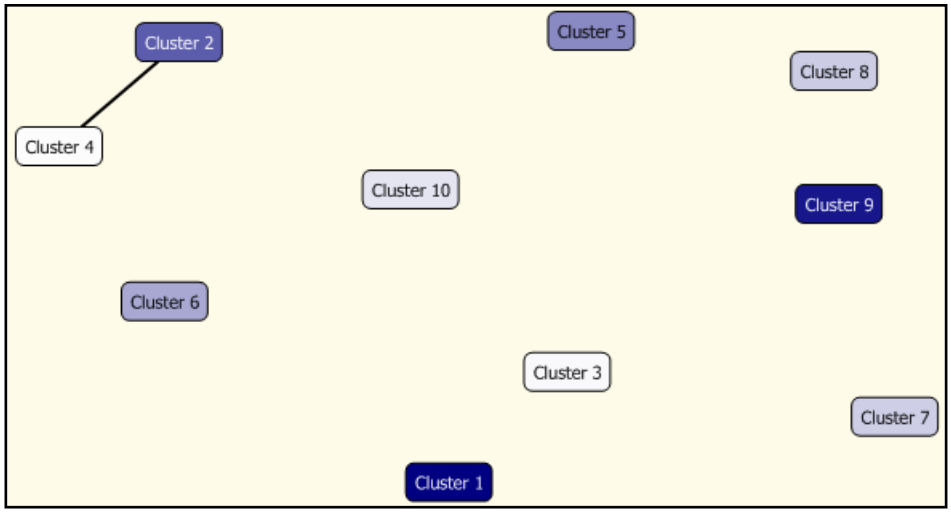
Uporabili bomo naslednje tehnike:
- Microsoft Clustering

Podatki se nahajajo v tabelah:
- Postavka kupca
- Kupec
- Podrobna postavka kupca

Trajanje podatkovnega rudarjenja:
Osvežitev podatkov: 10min 3s

4.1.2.1 Pridobljeno znanje

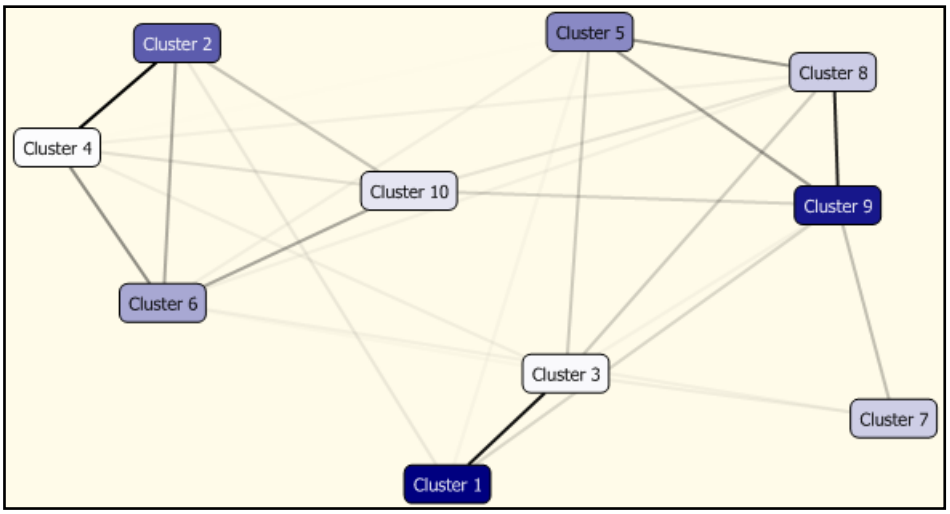
Rezultat, ki ga pridobimo z zgoraj navedenimi tehnikami (na naših nekoliko spremenjenih podatkih), je zapisan spodaj.



Slika 4.2. Skupine strank, ki nakupujejo pri nas.

Komentar k sliki:

Prikazana povezanost je povezanost po dimenziji »Globalna dimenzija 2«. Na sliki je prikazana najmočnejša povezanost med skupinami, zato je vidna le ena povezava.
Iz slike lahko vidimo, da je največja povezanost med skupinama »Cluster 2« in »Cluster 4« (zato ju Microsoft Analysis Services poveže skupaj). Ti dve skupini sta si po nakupovalnih navadah najbolj sorodni.



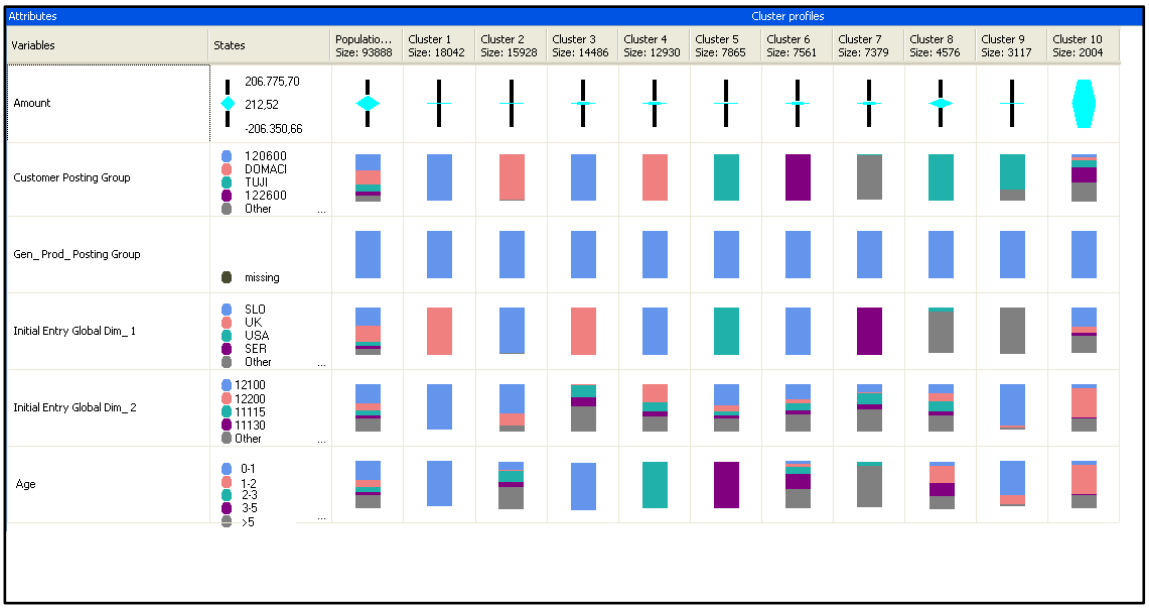
Slika 4.3. Skupine strank, ki nakupujejo pri nas.

Komentar k sliki:

Prikazana povezanost je povezanost po dimenziji »Globalna dimenzija 2«.
V primerjavi s prejšnjo sliko ta slika prikazuje nekoliko manjšo povezanost med skupinami. Algoritem je našel podobnosti (oz. »enakosti«) v obnašanju obeh skupin tudi med naslednjimi skupinami (v seznamu so urejene po padajoči moči povezovanja):

- Skupina 2 in Skupina 4,
- Skupina 1 in Skupina 3,
- Skupina 8 in Skupina 9,
- Skupina 4 in Skupina 6,
- Skupina 8 in Skupina 9,
- Skupina 5 je zelo dobro povezana s Skupino 8 in Skupino 9,
- Skupina 6 je zelo dobro povezana s Skupino 2 in Skupino 10,
- Skupina 2 in Skupina 10,
- Skupina 7 in Skupina 9,
- Skupina 3 je povezana s Skupino 5 in Skupino 8,
- Skupina 10 je povezana s Skupino 4 in Skupino 9,
- Skupina 1 in Skupina 9,
- Skupina 1 in Skupina 2,
- Skupina 4 je povezana s Skupino 5 in Skupino 7.

Iz zgornje slike lahko bralec vidi, da tudi moč povezave med skupinami, in sicer s pomočjo odtenkov črne barve povezav med vozlišči. Prikazane so vse povezave. Najmočnejša povezava je prikazana z najtemnejšo barvo. Prve tri skupine so zelo povezane med seboj, ostale manj.



Slika 4.4. Prikazuje razpored vrednosti po posameznih dimenzijah, ki sem jih določil kot zanimive.

Komentar k sliki:

Prikazane so 4 največjkrat prisotne vrednosti po posameznih dimenzijah. Tako lahko vidimo, da pri dimenziji »Customer Posting Group« pri Skupini 1 najbolj izstopa vrednost 120600.
Za Skupino 1 so očitno značilne vrednosti sledeče:
Customer Posting Group = '120600'
Gen_ Prod_ Posting Group = ''
Initial Entry Global Dim_ 1 = 'SLO'
Initial Entry Global Dim_ 2 = '12100'
Age = '0-1'

Iz zadnjega podatka lahko vidimo, da so v Skupini 1 stranke, ki so stare med 0 in 1 leti. Glede na to, da otrok med 0 in 1 letom starosti ne more kupovati, je očitno, da gre za podjetja, ki so nastala pred maksimalno 1 letom. Za nakupe pri nas se torej odločajo predvsem novonastala podjetja, kar je zanimiv podatek za segmentacijo strank.
Pri dimenziji Gen_ Prod_ Posting Group sem opazil, da so vrednosti pri vseh skupinah enako. Ne samo to, razvidno je, da uporabniki vrednosti pri nobeni transakciji niso vpisali. Na tak način lahko odkrijemo tudi napačne oz. nepopolne podatke.

4.1.2.2 Uporabnost Microsoft Clustering algoritma

Uporaba Microsoft Clustering algoritma je pomembna za razvrstitev strank v skupine. Omogoča nam, da bolje spoznamo nakupovalne navade naših strank. Če poznamo navade strank, se jim lahko tudi bolje prilagodimo. Opazimo lahko tudi določene nepravilne oz. nepopolne podatke.

Uporabna je tudi za ciljani marketing, saj nam omogoča, da kupce razvrstimo na tiste, ki kupujejo določeno skupino izdelkov, npr. skupino izdelkov 12900 in na tiste, ki kupujejo drugo skupino izdelkov, kupujejo npr. skupino izdelkov 12200. Tako lahko strankam pošiljamo različne ponudbe.

4.1.3 Napovedovanje prodaje

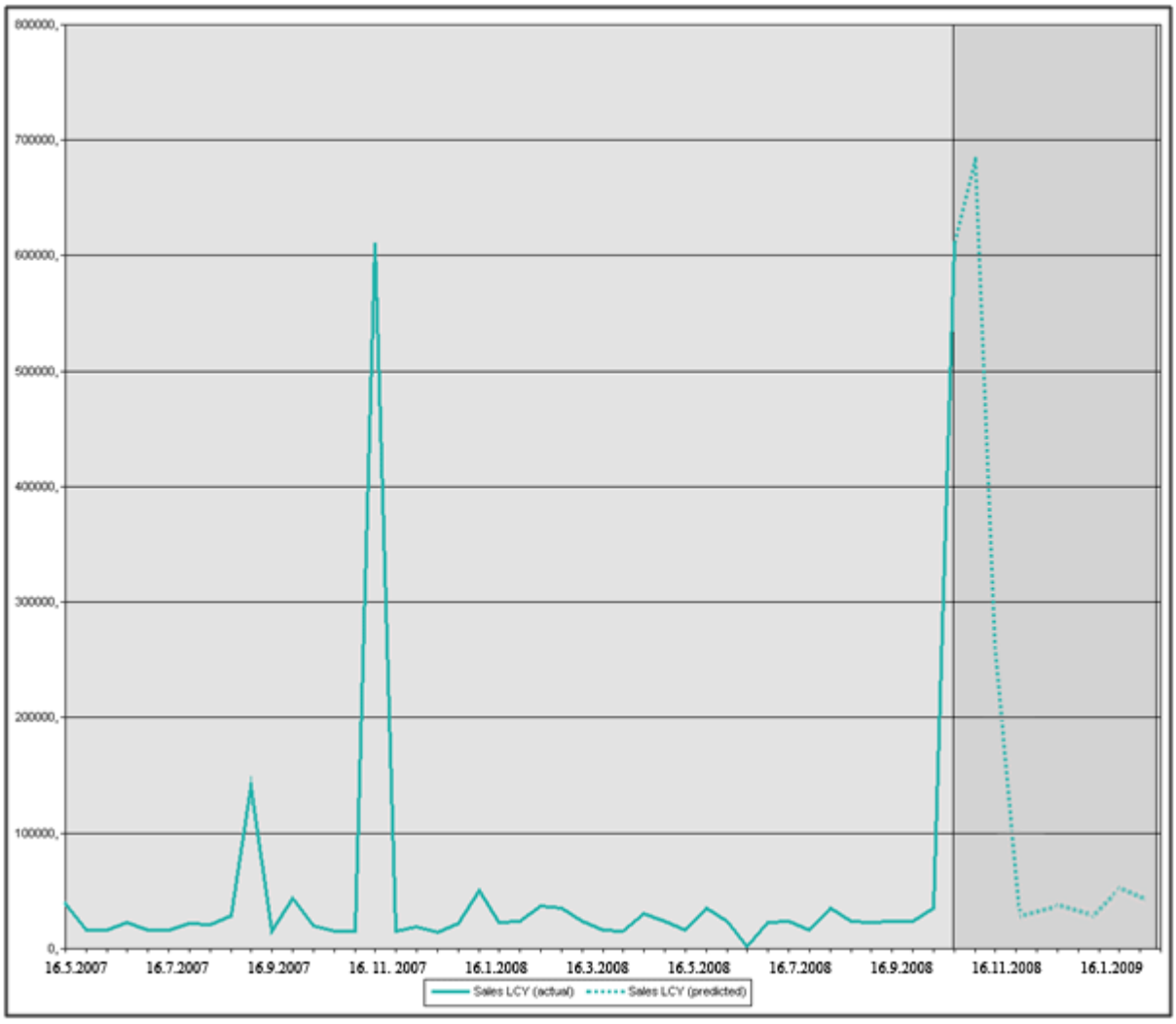
Uporabili bomo naslednje tehnike:
- Microsoft Time Series Algorithm (1),
- Microsoft Decision Tree (2).

Podatki se nahajajo v tabelah:
- Postavka kupca

Trajanje podatkovnega rudarjenja:
Osvežitev podatkov: 15min 12s

4.1.3.1 Pridobljeno znanje (1)

Rezultat, ki ga pridobimo z zgoraj navedenimi tehnikami (na naših nekoliko spremenjenih podatkih), je zapisan spodaj.

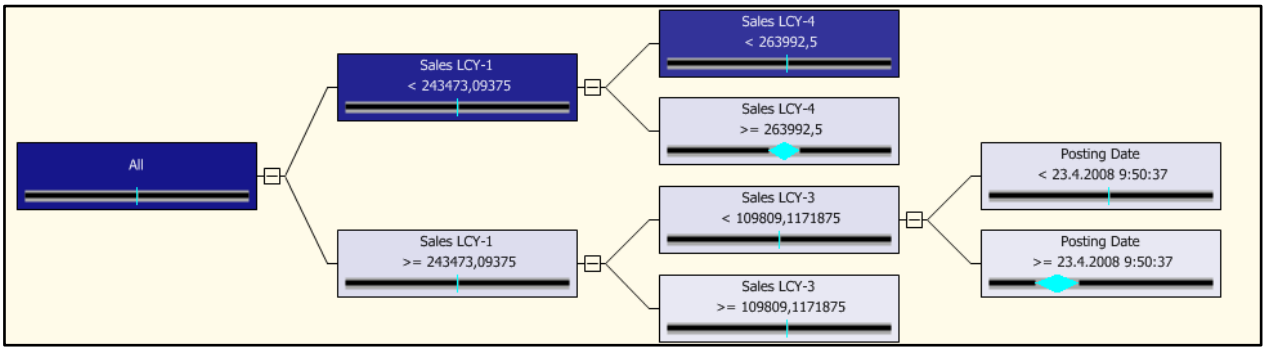


Slika 4.5. Prikazuje napoved za naslednjih nekaj mesecev.

Komentar k sliki:

Zaradi problematičnih podatkov (Datum knjiženja je v Microsoft Dynamics NAV v obliki "DD.MM.LLLL hh:mm:ss" – v relacijski bazi nimamo podatka samo o datumu) in premajhnega števila podatkov je prikazano samo nekaj naslednjih mesecev.
Najvišjo točko graf doseže pri Sales LCY = 682691,55549151963, ki je tudi napovedana vrednost prodaje, ki jo iščemo. To lahko vidimo na zavihku »Mining Model Prediction«.
Običajno podjetja napovedujejo prihodnje poslovne rezultate na podlagi prejšnjih let. Microsoft Analysis Services je glede na isto obdobje prejšnje leto (16. 11. 2007), ko je bila špica prodaje, letos (okrog datuma 16. 11. 2008) napovedal približno enako vrednost.

Poleg zgornje slike nam Microsoft Analysis Services prikaže še odločitveno drevo za zgornji graf, ki upošteva predvsem prodajo v predhodnjih obdobjih. Na zadnjem nivoju pa upošteva še Datum knjiženja.



Slika 4.6. Prikazuje odločitveno drevo za problem napovedovanja prodaje.

Microsoft Analysis Services je tako našel v podatkih naslednje zakonitosti:

Št.	Pogoj	Izračun
1	Sales LCY-1 < 243473,09375 and Sales LCY-4 < 263992,5	Sales LCY = 244446,464314045
2	Sales LCY-1 < 243473,09375 and Sales LCY-4 >= 263992,5	Sales LCY = 155958,399979764 - 0,00157302790733874 * Sales LCY(-4) - 0,00143174129104705 * Sales LCY(-3) - 0,746018993551464 * Sales LCY(-1)
3	Sales LCY-1 >= 243473,09375 and Sales LCY-3 < 109809,1171875 and Posting Date < 23.4.2008 9:50:37	Sales LCY = 244446,464314045
4	Sales LCY-1 >= 243473,09375 and Sales LCY-3 < 109809,1171875 and Posting Date >= 23.4.2008 9:50:37	Sales LCY = -1253600,54394262 - 0,365263139371729 * Sales LCY(-1) + 115,866467545968 * Sales LCY(-4) - 23,5893202801311 * Sales LCY(-3)
5	Sales LCY-1 >= 243473,09375 and Sales LCY-3 >= 109809,1171875	Sales LCY = 244446,464314045

Tabela 4.2. Rezultati za računanje Sales LCY glede na prehodni Sales LCY.

Komentar k tabeli izračunov:

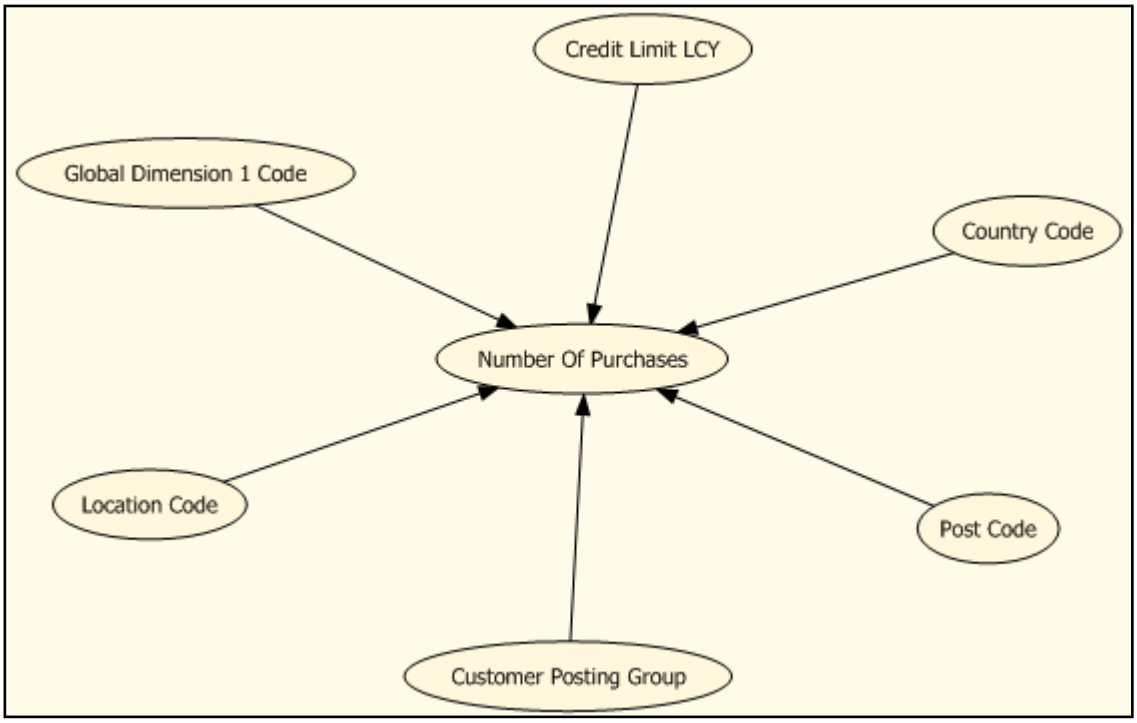
Zgornja tabela nam v povezavi z odločitvenim drevesom prikaže prihodnje vrednosti prodaje. Nekoliko čudno se mi zdi, da pri točkah 1, 3 in 5 algoritem napove iste številke, kljub različnim pogojem (gleda samo prodajo v predhodnjih obdobjih in ne Datuma knjiženja). Možna vzroka za ta odstopanja so lahko:

- Sezonska gibanja (npr. podjetje pospešuje prodajo vedno decembra, ker želi zadostiti napovedim),
- Odprtje podjetja (začetek prodaje, ko imamo še nekoliko nepopolne podatke).

Poleg zgornjih dveh razlogov pa je lahko razlog tudi ta, da preprosto ni vzorca v podatkih in so zato rezultati nelogični.
Na podlagi enega testa in enega nabora podatkov je težko z zadostno verjetnostjo ugotoviti kaj je razlog. Potrebovali bi večkratne teste, pomembno pa je tudi, da imamo testno množico reprezentativno in hkrati ločeno od učne množice.

4.1.3.2 Pridobljeno znanje (2)

Rezultat, ki ga pridobimo z zgoraj navedenimi tehnikami (na naših nekoliko spremenjenih podatkih), je zapisan spodaj.



Slika 4.7. Odvisnost števila nakupov od ostalih parametrov.

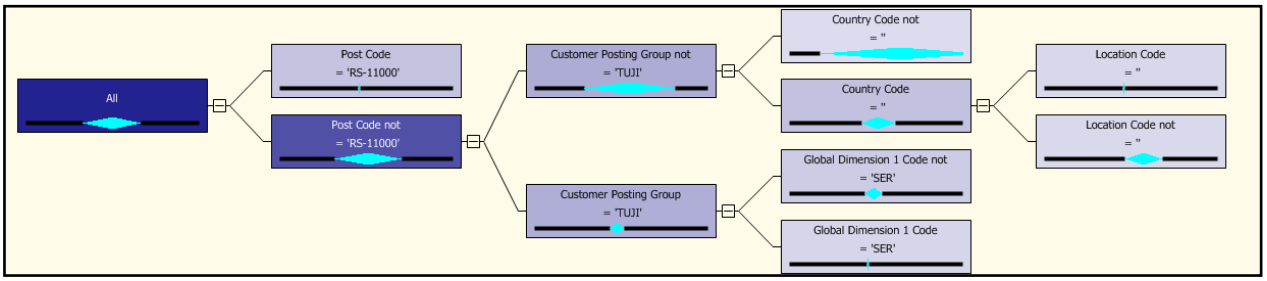
Komentar k sliki:

Število nakupov je odvisno predvsem od vrednosti treh dimenzij (po padajoči pomembnosti):

- Post Code, ki predstavlja poštno številko stranke,

- Customer Posting Group, ki predstavlja knjižno skupino stranke,
- Global Dimension 1 Code, ki predstavlja trg, kjer je stranka kupila izdelek,
- Country Code, ki predstavlja državo stranke,
- Location Code, ki predstavlja kodo lokacije, kjer je stranka kupila izdelek,
- Credit Limit LCY, ki pove, kolikšen limit ima stranka pri nakupih.

Microsoft Analysis Services nam prikaže odločitveno drevo , ki upošteva predvsem vrednosti v poljih »Post Code«, »Customer Posting Group«, »Global Dimension 1 Code«, »Country Code« in »Location Code«. Pri izračunu lojalnosti strank pa upošteva še vrednost polja »Credit Limit LCY«. Ostala polja imajo vrednosti, v katerih Microsoft Analysis Services ne najde vzorcev v podatkih.



Slika 4.8. Prikazuje odločitveno drevo za problem napovedovanja števila nakupov.

Microsoft Analysis Services je tako našel v podatkih naslednje zakonitosti:

Št.	Pogoj	Izračun	Povprečje
1	Post Code = 'RS-11000'	Number Of Purchases = 5,036+0,0000003*(Credit Limit LCY-3.571.428,571)	3,9633
2	Post Code not = 'RS-11000' and Customer Posting Group not = 'TUJI' and Country Code not = ''	Number Of Purchases = 1.908,727-2,041*(Credit Limit LCY-90,909)	2094,2313
3	Post Code not = 'RS-11000' and Customer Posting Group not = 'TUJI' and Country Code = '' and Location Code = ''	Number Of Purchases = 26,933	26,9333
4	Post Code not = 'RS-11000' and Customer Posting Group not = 'TUJI' and Country Code = '' and Location Code not = ''	Number Of Purchases = 1.224,214+0,003*(Credit Limit LCY-236.626,643)	574,3135
5	Post Code not = 'RS-11000' and Customer Posting Group = 'TUJI' and Global Dimension 1 Code not = 'SER'	Number Of Purchases = 322,391+0,002*(Credit Limit LCY-143.478,261)	45,5803
6	Post Code not = 'RS-11000' and Customer Posting Group = 'TUJI' and Global Dimension 1 Code = 'SER'	Number Of Purchases = 2,313	2,3132

Tabela 4.3. Rezultati za računanje števila nakupov.

Komentar k tabeli izračunov:
Število nakupov je odvisno od tega koliko kredita določimo stranki (»Credit Limit LCY«), na podlagi tega se izračuna število nakupov.
Največje povprečje nakupov imajo stranke, ki v polju »Customer Posting Group« nimajo vrednosti 'TUJI', imajo vrednosti v polju »Country Code« in »Post Code« ni 'RS-11000'.
Najmanjše povprečje nakupov pa stranke, ki prihajajo iz Republike Srbije, »Post Code« ni 'RS-11000', »Customer Posting Group« pa je enak 'TUJI'.

4.1.3.3 Uporabnost Microsoft Time Series algoritma

Na enak način lahko pristopimo tudi pri napovedovanju razlike v ceni, količine, ... Odvisno je le od količine in kakovosti podatkov, ki jih shranimo v podatkovno bazo.

4.1.3.4 Uporabnost Microsoft Decision Tree algoritma

Tu sem naredil napovedovanje števila nakupov na podlagi različnih vrednosti polj, kar nam lahko pomaga pri delitvi strank na enkratne in večkratne. Večkratne stranke je treba še dodatno nagrajevati za njihovo zvestobo, saj nam take stranke prinesejo največ dobička.

4.1.4 Ohranitev kupca

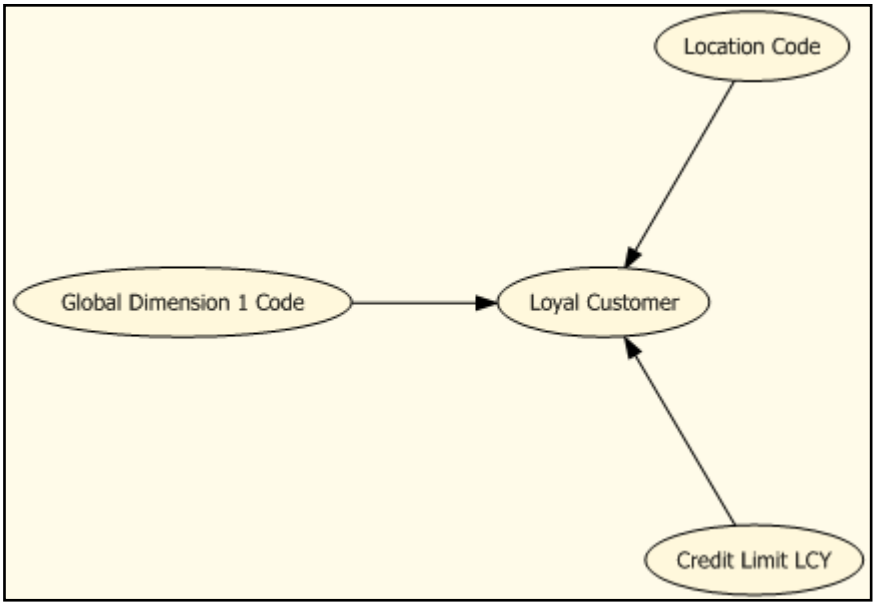
Uporabili bomo naslednje tehnike:
- Microsoft Decision Tree

Podatki se nahajajo v tabelah:
- Kupci,
- Postavka kupca,
- Podrobna postavka kupca.

Trajanje podatkovnega rudarjenja:
Osvežitev podatkov: 7min 33s

4.1.4.1 Pridobljeno znanje

Rezultat, ki ga pridobimo z zgoraj navedenimi tehnikami (na naših nekoliko spremenjenih podatkih), je zapisan spodaj.



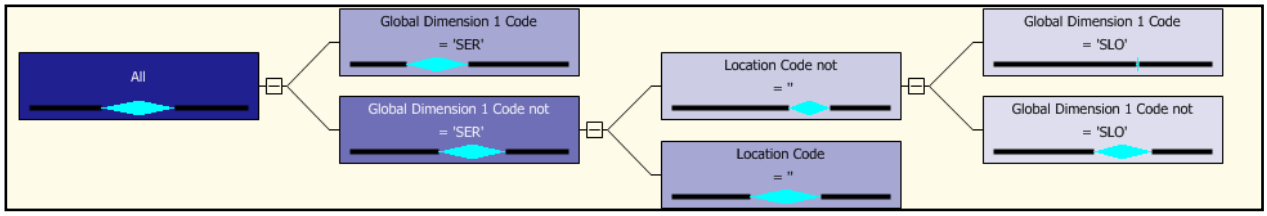
Slika 4.9. Odvisnost lojalnosti strank od treh parametrov.

Komentar k sliki:

Lojalnost strank je predvsem odvisna od vrednosti treh dimenzij (po padajoči pomembnosti):
- Global Dimension 1 Code, ki predstavlja trg, kjer je stranka kupila izdelek,
- Credit Limit LCY, ki pove, kolikšen limit ima stranka pri nakupih,
- Location Code, ki predstavlja kodo lokacije, kjer je stranka kupila izdelek.

Credit Limit LCY je pomemben predvsem zaradi izračuna same vrednosti lojalnosti strank (v odločitvenem drevesu se nahaja pod Izračunom). Lahko bi posplošili, da več kot damo stranki kredita, bolj zadovoljna in lojalna bo (s tem bo več denarja zapravila pri nas).

Microsoft Analysis Services nam prikaže odločitveno drevo, ki upošteva predvsem vrednosti v poljih »Global Dimension 1 Code« in »Location Code«. Pri izračunu lojalnosti strank pa upošteva še vrednost polja »Credit Limit LCY«.



Slika 4.10. Prikazuje vpliv različnih dimenzij na lojalnost stranke.

Microsoft Analysis Services je tako našel v podatkih naslednje zakonitosti:

Št.	Pogoj	Izračun
1	Global Dimension 1 Code = 'SER'	Loyal Customer = 0,238
2	Location Code not = '' and Global Dimension 1 Code = 'SLO'	Loyal Customer = 1,000
3	Global Dimension 1 Code not = 'SER' and Location Code not = '' and Global Dimension 1 Code not = 'SLO'	Loyal Customer = 0,800+0,0000008*(Credit Limit LCY-310.000,000)
4	Global Dimension 1 Code not = 'SER' and Location Code = ''	Loyal Customer = 0,595+0,000002*(Credit Limit LCY-4.785,714)

Tabela 4.4. Rezultati za napovedovanje lojalnosti kupca.

Komentar k tabeli izračunov:

Lojalnega kupca sem definiral kot kupca, ki je v zadnjem letu opravil vsaj en nakup. Tak kupec ima v polju vrednost Loyal Customer enako 1.
Zanimivo iz zgornjih podatkov je, da so kupci, pri katerih je »Global Dimension 1 Code« enako 'SLO' in nimajo izpolnjenega polja »Location Code«, vsi lojalni.
Srbski kupci pa so v povprečju lojalni zgolj v 23,8% primerov.

4.1.4.2 Uporabnost Microsoft Decision Tree algoritma

Opazil sem, da lahko na podlagi dobrih podatkov iz obstoječih zapisov v bazi s pomočjo Microsoft Decision Tree algoritma zgradimo model, ki nam lahko vnaprej pove kateri kupci so lojalni. To bi bilo dobro preveriti na večji množici podatkov, da bi lahko dokazali pravilnost delovanja.

5 Sklep

5.1 Zaključek

Pogoj za podatkovno rudarjenje je, da ima podjetje veliko količino podatkov v svojih podatkovnih bazah in da so ti podatki kvalitetni. Podatkovno rudarjenje namreč potrebuje veliko podatkov, da se lahko iz le-teh nekaj nauči oz. najde kakšne zakonitosti v njih; pravilnejši vzorci so razvidni pri večjih količinah podatkov.

Zaradi velike količine podatkov težko izluščimo koristne informacije iz podatkovnih baz, saj lahko zaradi današnje velikosti podatkovnih baz oz. skladišč kaj hitro kakšno zakonitost spregledamo.

Pozitivni učinki, ki jih pridobimo z vpeljavo podatkovnega rudarjenja, so sledeči:

- Večji prihranki (po začetnih sicer višjih denarnih vložkih),
- Boljše napovedovanje finančnih rezultatov,
- Boljše napovedovanje gibanja prodaje,
- Boljše razumevanje strank omogoča, da se stranki bolj posvetimo, kar je še posebno učinkovito v povezavi s CRM sistemi. Bolj kot stranko poznamo, lažje ji kaj prodamo. Več kot jih prodamo, večji so dobički.

Podatkovno rudarjenje je bolj primerno za večja podjetja z večjimi proračuni za vlaganje v informatiko. Za podatkovno rudarjenje velja, da je običajno dolgotrajen proces, kar pa za seboj prinese tudi večje stroške. Podatke moramo najprejočistiti, nato povezati različne vire v skupno skladišče. Običajno se naredi tudi podatkovno skladišče, ki je precej dolgotrajen proces in si ga lahko privoščijo le največja podjetja. Poleg izgradnje podatkovnega skladišča je tudi strojna oprema, na kateri tečejo omenjeni algoritmi za podatkovno rudarjenja, precej draga.

Podjetje se mora tudi odločiti za kaj točno bodo uporabljali podatkovno rudarjenje oz. nad katerimi podatki. Postaviti si mora prioritete; katere podatke bi radi najprej analizirali.

Kot dobro se izkaže, da naročnik izbere sponzorja projekta, ki žene projekt naprej in daje navodila in smernice za projekt, saj le to pri velikih projektih omogoča večjo možnost uspeha projekta. Ker je projekt podatkovnega rudarjenja zelo kompleksen projekt in zahteva veliko vsebinskega znanja, je dobro, da si izvajalec projekta skupaj z naročnikom vzame čas in prediskutira pridobljene informacije. Razumevanje znanja pridobljenega iz podatkov, je namreč ključnega pomena za uspešnost projektov podatkovnega rudarjenja. Poleg tega ni nujno, da podatkovno rudarjenje vedno najde koristne rezultate.

Kljub sorazmerno visokim stroškom vpeljave, se vpeljava še vedno (v večini primerov) obrestuje pri podjetjih, ki imajo opravka z veliko količino podatkov in želijo izvajati pravočasne in pravilne odločitve. Vse več podjetij se zaveda, da so podatki nekoristni, če jih neznamo pretvoriti v informacije oz. znanje.

Podatkovno rudarjenje je konkurenčna prednost podjetja pred tekmeci!

5.2 Problemi vpeljave DM

Probleme vpeljave podatkovnega rudarjenja lahko razdelimo na dva vidika:
- poslovni,
- tehnični.

Velik problem se kaže v sposobnostih organizacije za **pravilno razumevanje rezultatov procesa podatkovnega rudarjenja**. Več poslovnih faktorjev vpliva na končen rezultat vpeljave podatkovega rudarjenja v podjetje.

Vprašanje pravilnosti napovedovanja je vsekakor eno izmed vprašanj, ki ga je dobro vzeti pod drobnogled, saj napovedujemo poslovanje podjetja v prihodnosti na podlagi preteklosti. Tako napovedujemo prodajo na podlagi vedenja strank v preteklosti in ne moremo (vsaj ne z veliko natančnostjo) napovedati, kaj bomo prodali stranki v prihodnosti. Kljub temu, da izbire lahko razkrijejo želje strank, izbire in želje ne sovpadajo nujno. Želje so bolj trdno zasidrane v nas samih kot izbire, ampak običajno informacij o željah strank nimamo (razen, če je bila izvedena anketa prav s tem namenom!).

Običajen problem je tudi **pravilno vrednotenje novih informacij**, ki smo jih dobili z vpeljavo novosti v poslovno okolje. V iskanju resnično neznanih informacij morajo analitiki zniževati nivo zaupanja v rezultate podatkovnega rudarjenja. V takih situacijah pa se morajo analitiki vedno znova vprašati ali so te informacije sploh pomembne ali gre samo za enkraten pojav, do katerega pride pri spreminjanju parametrov pri podatkovnem rudarjenju.

Zelo pomemben faktor pri vpeljavi podatkovnega rudarjenja je **možnost integracije podatkovnega rudarjenja v obstoječe poslovno okolje** in aplikacije. Posledica tega je tudi verjetnost, da bodo uporabniki začeli uporabljati novo rešitev. Znano je, da se uporabniki različnih programov v podjetju z odporom učijo novih programov in jih velikokrat tudi (neupravičeno) zavrnejo [1, stran 129].

S tehničnega vidika pa na proces podatkovnega rudarjenja vplivajo predvsem:
- **Slaba kvaliteta podatkov**, ki jih hranimo v podatkovnih bazah, npr. so pomanjkljivi oz. napačno vpisani s strani ljudi (človeški faktor je še vedno najbolj pogost pojav napačnih podatkov! [6, strani 175-176, 232-234]). Lahko pa tudi pri izdelavi modela ne izberemo pravilnih podatkov za vhodne parametre.
- Podatki so velikokrat shranjeni **v kompleksnih podatkovnih strukturah**, saj so prvotno namenjeni za efektivnost programov, ki jih uporabljajo v podjetju, in ne toliko za podatkovno rudarjenje. Iz takšnih podatkov je pogosto težko narediti konsistenten in zanesljiv model za namene podatkovnega rudarjenja.
- **Težka dostopnost podatkov** iz različnih podatkovnih baz, kjer moramo pri podatkovnem rudarjenju pogosto povezovati različne vire podatkov.
- **Povečevanje količine različnih spremenljivk** pri procesu podatkovnega rudarjenja eksponentno vpliva na kompleksnost samega procesa [1, stran 129].

Velikokrat je **problem tudi v znanju in spretnostih**, ki jih ljudje potrebujejo za vodenje projektov podatkovnega rudarjenja (dobre spretnosti s področja podatkovnega rudarjenja so

redkost, še bolj redko pa jih najdemo v povezavi z vsebinskim znanjem iz gospodarstva in realnih poslovnih situacij).

Podatkovno rudarjenje prav tako nima določenega še nobenega standarda za primerjavo sposobnosti, performans in točnosti različnih modelov. To izhaja iz dejstva, da je podatkovno rudarjenje sorazmerno mlada disciplina.

6 PRILOGE

V tem poglavju se nahajajo razširjene tabele nekaterih prej omenjenih tabel.

Razširjeni zapisi iz zgoraj navedenih tabel:

Šifra kupca	Naziv	Naslov	ZIP	Kraj	Velikost podjetja*	Čas obstoja	Osnovni kapital	Dimenzija 1	Dimenzija 2	Knjižna skupina kupca	Koda prodajalca	Država	Datum prvega nakupa	Datum nastanka / rojstva
K1082612112	Kupec K1082126112	TOVARNIŠKA 12	5270	AJDOVŠČINA	mikro	2	8500	-	705	120010	10	SI	23.1.2007	15. 05. 2006
K1082126878	Kupec K1082126878	KALE 184	3320	VELENJE	veliko	71	112500	-	705	120000	5	SI	3.9.2007	01. 01. 1937
K0016125501	Kupec K0016125501	KOSOVELOVA 2	1000	GROSUPLJE	končni	25	-	-	705	121000	11	SI	3.4.2007	20. 03. 1983
K1016125502	Kupec K1016125502	HEINZELOVA 62 A	3300	ZAGREB	veliko	44	2500123	3035	191	165500	5	HR	3.4.2007	19. 09. 1964
K1016125507	Kupec K1016125507	HALILOVCI 12	3000	SARAJEVO	veliko	50	153688	3031	70	165500	23	BA	3.4.2007	22. 11. 1958
K1016125510	Kupec K1016125510	SLOVENSKA 55	6400	LJUBLJANA	mikro	13	12000	-	705	165500	23	SI	3.4.2007	26. 2. 1995
K1016125526	Kupec K1016125526	Prve Pile 1	5400	ZAGREB	veliko	35	3548995	-	191	165500	5	HR	3.4.2007	03. 01. 1973
K1016125529	Kupec K1016125529	KOLODVORSKA 7	3500	LJUBLJANA	končni	19	-	3013	705	121000	50	SI	3.4.2007	08. 08. 1989
K1016125530	Kupec K1016125530	Stegne 19	1400	LJUBLJANA	končni	53	-	3013	705	121000	12	SI	3.4.2007	03. 10. 1955
K1016125533	Kupec K1016125533	PRESERJE 60	6900	PRESERJE	mikro	9	11157	1026	705	165500	9	SI	3.4.2007	13. 09. 1999

Tabela 6.1. Razširjeni zapisi iz tabele Kupci.

* **Kategorizacija velikosti podjetja po ULRS [9]**:
- **veliko podjetje** je podjetje z več kot 250 zaposlenimi
- **srednje veliko podjetje** je podjetje, v katerem število zaposlenih ne presega 250
- **malo podjetje** je podjetje, v katerem število zaposlenih ne presega 50 ljudi
- **mikro podjetje** je podjetje, v katerem število zaposlenih ne presega 5
- če gre za končnega kupca piše »**končni**«

Poleg števila zaposlenih so v uradni klasifikaciji podjetij upoštevani tudi drugi kriteriji, predvsem kapitalska intenzivnost, vendar bomo v nadaljevanju največkrat navajali samo velikost in s tem razumeli uradno klasifikacijo

Št. postavke	Datum knjiženja	Tip postavke	Št. artikla	Količina	Šifra lokacije	Preostala količina	Zaračunana količina	Datum dokumenta	Dimenzija 1	Dimenzija 2
13	1. 3. 2008	Nakup	2930002	5	01	1	5	1. 3. 2008	3035	191
378	3. 3. 2008	Prodaja	2930002	-2	01	0	-2	1. 3. 2008	3035	191
501	19. 3. 2008	Prodaja	2930002	-2	01	0	-2	17. 3. 2008	3035	191
502	19. 3. 2008	Nakup	1152201	17	02	0	17	19. 3. 2008	3031	70
503	20. 3. 2008	Nakup	1355549	5	01	3	5	18. 3. 2008		705
504	20. 3. 2008	Prodaja	1152201	-3	02	0	-3	18. 3. 2008		705
505	20. 3. 2008	Prodaja	1355549	-2	01	0	-2	20. 3. 2008		705
506	20. 3. 2008	Prenos	1152201	-17	02	0	-17	20. 3. 2008	3031	70
507	20. 3. 2008	Prenos	1152201	17	01	17	17	20. 3. 2008	3031	70

Tabela 6.2. Razširjeni zapisi iz tabele Postavka artikla.

Št. postavke	Št. postavke artikla	Znesek stroška (dejanski)	Datum knjiženja	Strošek na enoto	Datum dokumenta	Šifra lokacije	Šifra dokumenta
22	13	12,45	1. 3. 2008	2,49	1. 3. 2008	01	DN_121208
51	13	-0,40	5. 3. 2008	-0,08	4. 3. 2008	01	POP_1222
125	502	335,12	19. 3. 2008	19,713	19. 3. 2008	02	DN_030208
133	503	179,00	20. 3. 2008	35,80	20. 3. 2008	01	NAB_170308
145	502	112,20	22. 3. 2008	6,60	22. 3. 2008	02	PREVOZ_0308

Tabela 6.3. Razširjeni zapisi iz tabele Postavka vrednosti.

Šifra artikla	Naziv	Knjižna skupina artikla	Strošek enote	Posreden strošek	Zadnji neposredni strošek	Točka ponovnega naročanja	Maksimalna zaloga	Količina ponovnega naročanja
2930002	Svinčnik	115300	0,53	0,02	0,52	25	100	25
1152201	Nalivno pero	115300	12,33	0,13	12,31	0	0	0
1355549	Radirka	115300	0,13	0,01	0,13	25	200	50
1253356	CD 25/1	115400	11,31	1,01	11,3	5	40	10
5356626	Pisalni blok	115300	1,49	0,11	1,49	40	100	20

Tabela 6.4. Razširjeni zapisi iz tabele Artikel.

Številka	Ime in priimek	Delovno mesto	Naslov	Kraj	ZIP	Datum zaposlitve	Datum rojstva
1	Delavec 1	SKLADIŠČNIK	PREGLOV TRG 2	LJUBLJANA	1000	1.1.2005	12.9.1959
2	Delavec 2	VODJA ODD.FINANC	BEBLERJEV TRG 6	LJUBLJANA	1000	1.1.2005	12.3.1981
10	Delavec 3	PRODAJNI REFERENT	PREŠERNOVA 33	LJUBLJANA	1000	15.2.2006	15.7.1976
11	Delavec 4	PRODAJNI REFERENT	BUKOVICA 43	VODICE	1217	15.5.2006	1.3.1966
12	Delavec 5	PRODAJNI REFERENT	DRAŽGOŠKA 7	KRANJ	4000	15.5.2006	11.9.1977
23	Delavec 6	PRODAJNI REFERENT	CELOVŠKA 189	LJUBLJANA	1000	1.1.2007	13.8.1986
50	Delavec 8	PRODAJNI REFERENT	KRAŠKA CESTA 70	DIVAČA	6215	15.7.2007	15.5.1969

Tabela 6.5. Razširjeni zapisi iz tabele Delavec.

7 LITERATURA

[1] Cabena, Hadjinian, Stadler, Verhees, Zanasi: Discovering Data Mining from concept to implementation, Upper Saddle River (New Jersey) : Prentice Hall PTR, 1998.

[2] Chung, H.M.: Dependability in data mining: a perspective from the cost of making decisions, ARES 2006.

[3] Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy: Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press 1996.

[4] Han, Jiawei: Data Mining Concepts & Techniques, Morgan Kaufmann, 2001.

[5] Mihelčič, Miran: Poslovne funkcije, Ljubljana, Fakulteta za računalništvo in informatiko, 2000.

[6] Pang-Ning, Tan: Introduction to Data Mining, Boston [etc.] : Pearson Addison Wesley, 2005.

[7] Starman, Daniel: Neposredno trženje, Delovno gradivo, zapiski predavanj, Ekonomska fakulteta v Ljubljani.

[8] Reason, James: Human Error, Cambridge University Press, Cambridge, 2000.

[9] Republika Slovenija, Uradni list RS, št. 40/2004 z dne 20. 4. 2004. 3.člen.

[10] (2008) Orodje QlikView, dostopno na: www.qlikview.com.

[11] (2006) Data mining: it makes sense, dostopno na: http://reader.feedshow.com/show_items-feed=9cc0f6723d0c9519363f0602902a4d17?page=1.

[12] (2005) Association Analysis, dostopno na: http://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf.

[13] (1996) Piatetsky-Shapiro, Gregory: Machine Learning, Data Mining, and Knowledge Discovery: An Introduction, dostopno na: http://www.kdnuggets.com/data_mining_course/dm1-introduction-ml-data-mining.ppt

[14] (1996) Vizualization & Data Mining, dostopno na: http://www.kdnuggets.com/data_mining_course/dm15-visualization-data-mining.ppt.

[15] (2005) AJPES, dostopno na: http://www.ajpes.si.

[16] (2008) Email marketing terms, dostopno na: http://emailmarketingpro.org/email-marketing-terms/.

[17] (2006) Saar-Tsechansky, Maytal: Data Mining, dostopno na: http://www.mccombs.utexas.edu/faculty/Maytal.Saar-Tsechansky/Teaching/OptionIIDataMining/Slides/Intro&Basics-session1.ppt.

[18] (2008) Microsoft SQL Server 2005: Analysis Services, dostopno na: http://www.microsoft.com/sql/technologies/analysis/overview.mspx .

[19] (2008) Vadnica za delo z Microsoft Analysis Services, dostopno na: http://msdn.microsoft.com/en-us/library/ms167167(SQL.90).aspx .

8 Seznam slik

9 Seznam tabel

Izjavljam, da sem diplomsko delo izdelal samostojno pod vodstvom mentorja (doc. dr. Marko Bajec). Izkazano pomoč drugih sodelavcev sem v celoti navedel v zahvali.

Miha Batič